ESSAYS ON THE ECONOMICS OF EDUCATION AND INEQUALITY

by

Arjun Jain

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Economics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2025

Date of final oral examination: 04/17/2025

The dissertation is approved by the following members of the Final Oral Committee:
Christopher R. Taber, Professor, Economics
Matthew J. Wiswall, Professor, Economics
Ashley Swanson, Associate Professor, Economics
Naoki Aizawa, Associate Professor, Economics
Marguerite Burns, Professor, Population Health Sciences

For those who believed in me.

ACKNOWLEDGMENTS

I am extremely grateful to my advisors Chris Taber and Matt Wiswall. Chris has known me since my time in UW-Madison's terminal master's program, where he guided me through my first foray into academic research; Matt later helped me take my master's thesis to the next level and provided helpful advise for each chapter of this dissertation. Mentorship from Chris and Matt over the years has molded me into the researcher I am today. Thanks also to Ashley Swanson and Naoki Aizawa for invaluable insight into the health world. I appreciate that Ashley has always paid so much attention to the details while at the same time keeping me from getting bogged down by the minutiae; I appreciate Naoki's constant encouragement and focus on the big picture. I thank Marguerite Burns for welcoming me onto her team at the UW-Madison Institute for Research on Poverty (IRP), which is where my idea for the job market paper came from. Though I have tried here, the truth is that I cannot adequately put into words how thankful I am for the advice, support, and encouragement from each of the wonderful people on my committee. Thank you to many other faculty at Wisconsin and beyond for their time and advice, including Md Moshi Ul Alam, Laura Dague, Jesse Gregory, Agustín Gutierrez, JF Houde, John Kennan, Corina Mommaerts, Martin O'Connell, Jack Porter, Fernanda Rojas Ampuero, and Jeff Smith.

I also thank several graduate students who have become some of my closest friends over the years and have made me feel like I was never alone on this journey. A huge thanks to Laura Boisten and Karla Hernández Romero, whose support during the job market kept me going. Along with Laura and Karla, I thank several other friends and office mates, including Jeremy Kirk, Cassi Kirkland, Anna Lukianova, and Shannon Sledz; each of them made the office a place I wanted to be (even if none of them could get me there on the weekend). I am grateful for Andrew Smith: the department may list me as Andrew's mentor on paper, but I have learned at least as much from him as he has from me. Andrew and his wife, Madi, have become like family for me, and I will always look fondly upon the bike rides and all-around good times we have had together. I am also thankful for several other amazing friends that have made me a better researcher or were there for me when times got tough, including Derek Christensen, Andrea Franco Hernández, Giselle Labrador Badía, Ben Murray, Felipe Parra Escobar, and Mitchell Valdés Bobes.

My inspiration for research in the economics of education and the initial reason I chose to pursue a PhD was the fulfilling - albeit brief - experience I had as an intern at the Wisconsin Department of Public Instruction (DPI). I am incredibly grateful for the opportunity that I was given there and thank the many amazing people I worked with, including Sam Bohrod, Emily Colo, Maria Dobbins, Derek Field, Rob Franke, Scott Jones, Amy Marsman, and Laura Pinsonneault.

Thank you also to all of the staff at the UW-Madison IRP, Wisconsin DPI, and North Carolina Education Research Data Center (NCERDC), who made my research using restricted-access data possible. A special thanks to Carl Frederick (at DPI and then IRP) and Kara Bonneau (at NCERDC); their help during data application processes got my projects off the ground and their quick responses to my many emails over the years kept my progress going.

Thank you to Becca George for invaluable support and seemingly endless patience during the job market, to Kim Grocholski for thoughtful and detailed answers to the multitude of questions I have had over the years, and to Elizabeth Foste for support while I was a master's student making the decision to pursue a PhD. Thanks also to many other staff at the UW-Madison Department of Economics, including Julie Anderson, Vicki Fugate, and Tammy Herbst-Koel. I am grateful for a number of fellowships and travel grants that I received from the UW-Madison Department of Economics and Graduate School. I also acknowledge and appreciate financial support from the Alice Gengler Wisconsin Graduate Fellowship.

Finally, I thank my family for their love and support. To my dad, thank you for always being there for me no matter how tough things got; your encouragement during the job market meant the world to me. To my mom, thank you for wanting to hear about my research and thinking critically about how answers to my questions might solve problems that you have observed firsthand. To both of my parents, thanks for taking the time to apply your medical expertise to the health side of my job market paper. And to my grandma, thank you for always welcoming me into your home with open arms when I needed a breath from research or a break from Madison.

Disclaimers: The work in this dissertation has been made possible by confidential data from the Wisconsin Department of Public Instruction, the Wisconsin Department of Health Services, the North Carolina Education Research Data Center, and the North Carolina Department of Public Instruction. The authors thank each of the data providing agencies for the use of data for these analyses, and acknowledge that the data providing agencies do not certify the accuracy of the analyses presented. The conclusions of this research and the views expressed herein are those of the authors and not those of the data providing agencies.

ABSTRACT

This dissertation includes three essays on the economics of education, with a special focus given to inequalities and inequities across students from different subgroups. In the first chapter, joint with Jeremy Kirk, we generalize the canonical model of school value-added to identify heterogeneity in school effectiveness driven by differences in student health. To do this, we first match the universe of K-12 public school records for the state of Wisconsin to individual-level Medicaid claims and enrollment; we then measure student health with a random forest, predicting students' absence rates directly from observed diagnosis codes. We find that the dispersion in school effectiveness is up to 31% larger among unhealthy students than among healthy students, indicating that schools are more influential in determining academic outcomes among unhealthy students. We then investigate how school nurses and homebound teachers (those qualified to provide at-home instruction to severely ill or disabled students) improve student outcomes. While nurses have little effect on test scores, homebound teachers boost test scores of unhealthy students and close the achievement gap between healthy and unhealthy students by 16%.

In the second chapter, I investigate how a high school's curriculum - in particular, the array of advanced courses it offers - influences its quality, and how participation in a high school's advanced courses affects achievement gaps between students from low- and high-resource households. Equipped with student-level data from the North Carolina Education Research Data Center, I use a value-added model to estimate high school quality, which I decompose into factors from the curriculum, teacher characteristics, and student body composition. I find that the student composition is most closely tied to school quality and that little variation can be attributed to the curriculum or teacher characteristics. Turning to the curriculum more directly, while schools' advanced course offerings themselves have little direct impact on ACT scores or achievement gaps, the returns to participation in advanced courses are substantial. I also find that the students that select into advanced courses – those with higher prior ability and those from high-resource households – are the ones who experience the largest benefit of participation.

In the third chapter, I explore one-to-one technology initiatives, policies in which school districts provide computers to every student. Using a two-way fixed effects approach, I evaluate these policies' impacts on test scores and achievement gaps between students from low- and high-income households. To do this, I combine student-level public school records with hand-collected computer distribution plans for several of Wisconsin's largest public school districts. I find that one-to-one initiatives widen English Language Arts (ELA) achievement gaps by 0.03 standard deviations, driven by negative effects on students from low-income households. This could be due to a combination of technological dependence of ELA curricula and limitations of internet at home for students from low-income households. Despite negative short-term effects, I find suggestive evidence that these initiatives helped mitigate some of the widespread learning losses that hit school districts across the country during the COVID-19 pandemic.

1 BETTER SCHOOLS OR HEALTHIER STUDENTS? UNCOVERING HIDDEN HETEROGENEITY IN SCHOOL EFFECTIVENESS

1.1 Introduction

Previous research has found that teachers and schools generate vital inputs to the production of human capital during childhood (e.g., Chetty et al., 2014a; Walters, 2015), which in turn is a key determinant of a wide array of outcomes in adulthood, such as college attendance and earnings (e.g., Chetty et al., 2014b; Deming et al., 2014). Similarly, investment into improving health in childhood has been linked closely to longer-term outcomes such as health in adulthood, earnings, and socioeconomic status (e.g., Grossman, 1972; Case et al., 2005; Currie, 2009). Although educational inputs and child health may be related to each other in complex ways - jointly affecting long-term human capital, health, and socioeconomic outcomes - much of the literature has studied one or the other in isolation.

While a large literature has found lasting effects of poor neonatal and childhood health on long-term schooling outcomes (e.g., Case et al., 2005; Currie et al., 2010), few studies have identified the shorter-term mechanisms driving long-term effects.[1] Figlio et al. (2014) and Almond et al. (2018) point out that research into how poor health in childhood affects contemporaneous academic outcomes - something we aim to do in this paper - is important in (1) understanding effects on long-term outcomes, and (2) identifying how policy can improve long-term outcomes. Figlio et al. (2014) find the lack of literature studying the impacts of poor health in childhood on human capital during key developmental years to be a direct consequence of data limitations: studies that link childhood health to adult outcomes rarely have access to schooling data. Moreover, detailed education data rarely includes comprehensive information on student health.

In this paper we ask the following question: how do school inputs to the production of human capital differ based on student health? We overcome data challenges by matching the universe of K-12 public school records for the state of Wisconsin to individual-level administrative Medicaid claims and enrollment, creating a database of nearly two million student-year observations that include detailed health records.[2] Using a rich set of diagnosis codes, demographic characteristics, and academic outcomes, we are able to explore the interaction between health and educational inputs to understand (1) the short-term impacts of poor health in childhood, and (2) the ways in which schools can mitigate the negative effects of poor health. The importance of policy geared toward unhealthy students is highlighted by a 0.4 standard deviation (SD) achievement gap between healthy and unhealthy students.[3] We evaluate how the employment of homebound teachers (those

[1]Bharadwaj et al. (2018) and Figlio et al. (2014) are recent exceptions that investigate how poor neonatal health affects the path of cognitive development during childhood.

[2]In an agreement between Wisconsin's Department of Health Services, Department of Public Instruction, and the University of Wisconsin-Madison's Institute for Research on Poverty (IRP), the state agencies have provided extensive data to the IRP-maintained Wisconsin Administrative Data Core.

[3]This is a lower bound because it only includes students from lower-income households. If one were to supplement

qualified to provide at-home instruction to students that are disabled or severely ill)[4] and school nurses impact academic outcomes among less healthy students.

School effectiveness - commonly estimated through value-added frameworks - is one of the primary ways in which researchers measure the productivity of school inputs.[5] While value-added methods have been useful in informing policy (e.g., Currie and Thomas, 2000; Deming, 2009; Hanushek, 2009; Chetty et al., 2014a), the past literature has omitted student health, a key source of student heterogeneity that may in part be responsible for selection of students into schools. Furthermore, health may explain often-overlooked group-level differences in school effectiveness. Identifying the role of health may therefore be consequential for education policy: in the case of a student that is struggling in school as a result of poor health, should a policymaker focus intervention on teaching or should they target health instead? Understanding the types of policies that boost achievement among unhealthy students will have a direct effect on the achievement gap between healthy and unhealthy students: a policy closes the gap if its effect on unhealthy students exceeds its effect on healthy students.[6]

We find that health characteristics explain at least 6% of the variation in the Black-White achievement gap across schools, though there is a substantially larger overlap between the two achievement gaps due to correlations between health and race.[7] This indicates that school policy boosting unhealthy students' outcomes has the potential to close other longstanding achievement gaps that reach beyond health. Moreover, an investigation of student health in the context of education gives individual schools concrete strategies for boosting outcomes among struggling students, such as hiring nurses and homebound teachers. Looking only at demographic characteristics like race and income, school-side policy recommendations are less clear and often require large-scale district-level intervention (e.g., school choice) or government intervention (e.g., social workers).

In order to identify the different impacts schools have on students of varying health statuses, we first need to measure student health. This is challenging given the immense amount of information contained in our data: we observe over 40 thousand unique diagnosis codes across 25 million Medicaid claims. To deal with large datasets containing tens of thousands of sparsely populated variables, machine learning techniques have recently gained ground in the health literature (Obermeyer and Emanuel, 2016; Shatte et al., 2019);[8] the random forest is particularly well-suited to our Medicaid claims data, due to its success in efficiently handling large numbers of predictors, combining them in complex nonlinear and interactive ways (Kleinberg et al., 2015). We use a ran-

our Medicaid claims data with private insurance claims, the gap would likely be far larger.

[4]Qualification requires a bachelor's degree and special education teaching certification.

[5]Following the terminology developed by Rothstein (2006) and Abdulkadiroğlu et al. (2020), we define an effective school as one that generates above-average causal gains to students' test scores.

[6]If, in addition, the effect on healthy students is at least zero, the policy is a (subgroup-level) Pareto improvement on the status quo: healthy students are no worse off while unhealthy students are better off.

[7]We find that the Black-White gap among students from lower-income households hovers around 0.7 SD.

[8]In segmenting populations on the basis of health, several studies have used the Johns Hopkins ACG (Adjusted Clinical Group) Grouper System. Currie et al. (2010), however, find that the ACG Grouper System excludes some of the most common codes among children, such as asthma and ADHD, making this method less appealing for our context that involves a student population.

dom forest to predict students' health-related absence rates directly from diagnosis codes observed in their Medicaid claims. These predictions constitute our constructed health index, which we discretize for the purpose of distinguishing healthy and unhealthy students from each other. Using absence rates as the outcome is a simple extension of models used previously in health literature, but it provides an important bridge for empirical analyses that study how academic outputs are affected by health inputs. Our method is also remarkably general: by using different outcomes (emergency department visits, probability of death, etc.) one could adapt our method to a number of different contexts and populations, and by combining multiple outcomes one could construct a multidimensional health index.

To allow for health-based heterogeneity in school effectiveness, we introduce our measure of student health into a generalization of the canonical model of school value-added specified in Abdulkadiroğlu et al. (2020).[9] While a large economic literature has used value-added models to measure teacher and school effectiveness (e.g., Wright et al., 1997; Rothstein, 2010; Chetty et al., 2014a; Angrist et al., 2017; Abdulkadiroğlu et al., 2020; Gilraine et al., 2023), none have accounted for student heterogeneity in health. We provide evidence that value-added estimates derived from models that rely on the selection-on-observables assumption but omit health are biased upward in magnitude due to the systematic confounding of school effectiveness with student selection into schools based on health. This in turn results in upward bias in the dispersion in school effectiveness, which we measure to be as large as 5%.[10]

With our specification of value-added that interacts a school fixed effect with a health group fixed effect, we uncover key heterogeneity in school effectiveness that is missed by prior literature that focuses only on the student body as a whole without any concern for subgroups. We find that the dispersion in school effects is up to 31% larger among unhealthy students than among healthy students.[11] In other words, the difference between effective and ineffective schools is greater for unhealthy students, indicating that schools are more influential in determining outcomes among unhealthy students. One explanation for this result is that a wider range of school policies impact unhealthy students than healthy students: teaching-based interventions may boost outcomes among healthy students while teaching-based interventions *and* health-based interventions may boost outcomes among unhealthy students. Furthermore, much of the cross-school variation in value-added differences across healthy and unhealthy students is explained by the effect on the unhealthy students; this finding implies that school policy is an important way to boost outcomes among

[9]The general specification of value-added has no interaction with the school fixed effect (e.g., Chetty et al., 2014a; Abdulkadiroğlu et al., 2020), except in some cases a year fixed effect to allow time-variation if there are many years of data available (e.g., Gilraine et al., 2023).

[10]This may not be a concern in studies with adequate instruments for school enrollment or studies that use a control function approach that relaxes the selection-on-observables assumption. In fact, Abdulkadiroğlu et al. (2020) find that the standard OLS estimates are larger in magnitude than estimates from the control function approach, leading to dispersion that is inflated by as much as 25%.

[11]With English Language Arts scores as the value-added outcome, dispersion is 31% larger among unhealthy students than healthy students in elementary schools, 6% larger in middle schools, and 20% larger in high schools. With Math test scores, dispersion is 17% larger among unhealthy students in elementary schools and 24% larger in high schools; in middle schools, dispersion is comparable across health groups.

unhealthy students and close the achievement gap between healthy and unhealthy students.

Our specification also allows us to speak to "match effects" between students and schools in a more general way than approaches taken in prior literature.[12] Often, research has documented average match effects in the population rather than effects that vary from school to school (or teacher to teacher): Gershenson et al. (2022) find that certain teachers are more effective at teaching students of the same race while Steinberg and Garrett (2016) and Aucejo et al. (2022) find that some teachers are better for higher-achieving students. Some recent exceptions have allowed for more flexibility in match effects (Abdulkadiroğlu et al. (2020) allow for matches at the school level while Ahn et al. (2023) allow for matches on several dimensions at the teacher level), though no study has uncovered match effects based on student health, which we examine in this paper.

We estimate match effects based on student health at the school level; in other words, we are able to separately identify the gains to test scores that are generated among healthy students and unhealthy students *for each school*, rather than identifying only a specific *type of school* that is better-suited to students in one group or the other. By projecting these subgroup-level measures of school effectiveness on a vector of school characteristics, we identify a number of characteristics that are predictive of a better match rather than focusing on a single dimension. We find that in high schools, the fraction of the student body that is unhealthy is the largest driver of effectiveness, particularly among healthy students.[13] This suggests at least one of two things: (1) schools with more unhealthy students may divert more resources toward the unhealthy students but away from healthy students, and (2) policies that aim to improve student health could have positive spillover effects on the productivity of high school inputs.

Finally, we implement a two-way fixed effects design to understand the effects of two school policies - the hiring of nurses and homebound teachers - on a range of cognitive and non-cognitive student outcomes. While a large economic literature has investigated school policy (e.g., school vouchers, teacher reallocation, one-to-one initiatives), research on school nurses and especially homebound teachers is scant.[14] Our design follows de Chaisemartin and D'Haultfoeuille (2020), which importantly allows for differences in treatment timing and transitions *out of* treatment as well as into treatment. This will be important because we observe schools that cut down on nurses and homebound teachers as well as schools that hire new nurses and homebound teachers. The crucial assumption underlying the two-way fixed effects model is parallel trends in students' outcomes, which we find holds here. With no mandate on school nurses in Wisconsin, we observe roughly 75-90% of schools to have a nurse in any given year; homebound teachers are only found in around 5%-45% of schools in any given year (with most serving students in high schools), but there is much more variation over time. We find that nurses increase absence rates (likely because they send kids

[12] Much of the value-added literature has actually forgone a match effect investigation, measuring teacher or school effectiveness in less flexible ways such that they do not vary with student characteristics.

[13] We also find that in elementary schools, higher proportions of female teachers are associated with increased effectiveness, particularly among unhealthy students; in middle schools, effectiveness for both subgroups is positively correlated with teacher experience.

[14] Some health research has been done to understand how school nurses improve health outcomes (de Buhr et al., 2020), but little has been done regarding academic outcomes.

that feel unwell home), but have no effect on students' test scores. Homebound teachers, on the other hand, boost unhealthy high school students' Math scores by 0.065 SD while there is no effect on healthy students' scores.[15] With an estimated achievement gap of 0.4 SD between unhealthy and healthy students, we estimate that homebound teachers close the health-based achievement gap by 16%. Moreover, with health directly explaining at least 6% of the Black-White gap across schools as well as large overlap between health-based gaps and race-based gaps due to correlation between health and race, our findings imply that homebound teachers may have positive impacts that reach well beyond closing the health-based achievement gap.

The remainder of this paper is organized as follows. Section 1.2 discusses our data sources and key sample restrictions. Section 1.3 presents our random forest model for identifying student health and establishes the importance of including health in models of school value-added; Section 1.4 outlines the process of discretizing the health index to distinguish between healthy and unhealthy students. Section 1.5 first introduces our interacted specification of the value-added model that allows for health-based heterogeneity in school effectiveness and then presents the main results from the value-added estimation. Section 1.6 first sets up the two-way fixed effect design used to evaluate school policy and then presents the policy-relevant results. Section 1.7 discusses some implications of our findings and Section 1.8 concludes.

1.2 Data

Data Sources

We use data from the Wisconsin Departments of Public Instruction (DPI) and Health Services (DHS) that are linked based on an anonymized individual identifier. These datasets are part of the Wisconsin Administrative Data Core (WADC) that is maintained by UW-Madison's Institute for Research on Poverty (IRP). There are two key features of this data that make our paper possible. First, we have an individual-level link between Medicaid claims and enrollment data and K-12 public school records, a link that typically either does not exist or is quite expensive to obtain. Second, the statewide Medicaid claims and enrollment data contain detailed diagnosis codes from which we are able to identify child health. Derived from administrative claims panel data, we consider our health measures to be more reliable than those elicited from surveys for a number of reasons. Selective attrition that is common in longitudinal surveys is not an issue for us, and we have a much larger sample than is typically possible with survey data (Almond et al., 2018). In addition, measurement error in our case is likely less severe: while we still need to worry about children not going to the doctor when they're unhealthy, our physician-provided diagnosis codes are likely closer to true health than children's perceptions of their own health.

[15] The effect on unhealthy high school students' English Language Arts scores is 0.054 SD; the effect among healthy students is far smaller and imprecise. Effects are close to zero in elementary schools and larger but far less precise in middle schools due to the lower availability of homebound teachers.

The student-level school records date back to the 2005-06 academic year and run through the 2020-21 academic year.[16] Our primary outcome is standardized test scores, though we also observe attendance rate, discipline, dropout, grade progression, and high school completion. We observe the following demographic characteristics: student race and ethnicity (which is condensed into one variable in the data), sex, subsidized lunch eligibility, English language proficiency, and special education enrollment. For subsidized lunch eligibility, we separately identify eligibility for free lunch from eligibility for reduced-price lunch, which helps in better distinguishing between students within the lower-income population.

On the health side, we observe Medicaid enrollment as well as all claims for outpatient, inpatient, and emergency department visits dating back to January 2008 and running through 2020. Each Medicaid claim contains up to nine ICD-9 or ICD-10 diagnosis codes (International Classification of Diseases - 9th and 10th Revision).[17] The first (primary) code represents the reason medical care was sought and/or provided by the doctor; the remaining (secondary) codes typically represent other medical diagnoses someone received during a doctor's visit but for which they did not seek medical care. We use the monthly Medicaid enrollment data to identify students who are enrolled but make no claims; we posit these to be healthy individuals.

Lastly, we have DPI data describing school and district characteristics. Most importantly, we are interested in the employment of nurses and homebound teachers, due to the importance of these agents in identifying and helping unhealthy students. We also observe school-aggregate teacher characteristics on race/ethnicity, gender, experience, and education. We also have detailed information on school and district report cards; though the construction of these variables only started in 2014 (midway through our analysis period), they provide practical policy-relevant measures to which we can compare our estimates of school effectiveness. At the district level, we have several measures of spending per pupil that relate to different portions of the budget.

Background: Standardized Exams

There has been substantial change to the Wisconsin Student Assessment System (WSAS) since 2005. Prior to 2014, the Wisconsin Knowledge and Concepts Examination (WKCE) was administered to students in grades 3-8 and 10 every October. Changing academic standards resulted in a substantial alteration and expansion of the assessment system in 2014. That year, the Badger exam replaced the WKCE for grades 3-8, though it was quickly replaced by the Forward exam in 2015. The Forward remained the exam for grades 3-8 through 2020 (the end of our data). From 2014 onward, the ACT Aspire was administered to students in grades 9-10 while the ACT was

[16]For simplicity, we will refer to an academic year (which DPI calls a "school year") by the fall in which it starts; for example, 2005 represents the 2005-06 academic year. DPI lists the start of an academic year as July and the end of that academic year as the following June.

[17]Prior to October 2015, ICD-9 codes were used; ICD-10 codes have been used since. ICD-9 and -10 codes were designed by the World Health Organization and are "used by physicians to classify and code all diagnoses, symptoms, and procedures for claims processing," according to the American Medical Association.

administered to grade 11. All exams subsequent to the WKCE were administered every spring.[18] Alternate exams are provided for students with significant cognitive disabilities, though they are not included in our analysis because few students take them and the results from these exams are very difficult to compare with those from the standard exams. The exam subjects we focus on in this paper are English Language Arts (ELA)[19] and Mathematics.

Background: Subsidized Lunch Eligibility

Subsidized lunch eligibility[20] is determined prior to each academic year primarily based on household income. Children from families with income equal to or below 130% of the federal poverty line (FPL) are eligible for free lunch at school, while children from families with income between 130% and 185% FPL are eligible for reduced-price meals. The two primary methods for determining eligibility are direct certification and applications submitted prior to the start of the academic year. Direct certification automatically enrolls children in the subsidized lunch program if they participate in FoodShare (SNAP), W-2 cash benefits (TANF), or Food Distribution Programs on Indian Reservations (FDPIR). Students are also categorically eligible if they are determined to be homeless, runaway, migrant, foster, or enrolled in a Head Start program.

Background: Medicaid Eligibility

Medicaid is administered at the state level, but subject to broad federal standards and funded jointly by the federal government and the state. In Wisconsin, the BadgerCare Plus program began in 2008 to expand Medicaid coverage to all children without health insurance regardless of income.[21,22] Once enrolled, recipients must renew each year to maintain Medicaid coverage, a process known as renewal or recertification (Dague and Myerson, 2024). One consequence of the recertification process is coverage loss among those who did not renew but were eligible for Medicaid, which can lead to short-term gaps in coverage among those with no change in eligibility (Aizer, 2007).

Linking the Medicaid Data to the Student Records

Though we have the anonymized individual identifiers in both the Medicaid data and the K-12 public school records, there is some nuance involved in linking the two together. In the Medicaid claims and enrollment data, we construct a year variable that matches to the academic year of the next standardized exam in the DPI school records. To illustrate this, suppose that exam $e-1$ occurs

[18] DPI gives a March - April window for administration of these exams; for simplicity we denote April as the exam month for all schools in our analyses.

[19] English Language Arts - which replaced Reading in 2014 due to changing standards - includes several areas of reading and writing, and examines what is broadly referred to as English literacy.

[20] Throughout this paper the terms "eligible for subsidized lunch," and "economically disadvantaged" will be used interchangeably, as the latter is DPI's preferred terminology.

[21] From Wisconsin DHS (https://www.dhs.wisconsin.gov/badgercareplus/clawaiver-app.pdf).

[22] In years prior, children could be enrolled in Medicaid through either presumptive eligibility or household application. Under presumptive eligibility, medical care providers can provide Medicaid coverage at the time of service for children who were likely to qualify based on household income.

in academic year $t-1$ and exam e occurs in academic year t; we match all claims occurring between exams $e-1$ and e to the academic year t. This allows us to account for around a year's worth of diagnoses from the Medicaid claims leading up to the following standardized exam. In terms of enrollment, we denote a student as being enrolled in Medicaid during academic year t if they were enrolled for at least one month in that academic year.

Sample Restrictions

We start with the universe of Wisconsin's K-12 public school students in from 2005 through 2020. We drop the first three years of data (2005-2007) because we only observe Medicaid claims and enrollment beginning in 2008,[23] and we drop the last two years (2019-2020) because standardized tests for these years were disrupted by the pandemic; we are left with student records from 2008 through 2018. We note here that 2008 is included in the random forest for the purpose of lagged variables, but we exclude it from the value-added analysis because we are unable to estimate a health index for students in this year.

We identify student health from Medicaid claims and enrollment data; consequently, we cannot fully determine the health of students while they are not enrolled in Medicaid,[24] so we need to make some additional sample restrictions. Furthermore, while all children in Wisconsin became eligible for Medicaid in 2008, enrollment among children is closely tied to eligibility among parents (i.e., lower-income parents that qualify for and enroll in Medicaid also enroll their children); American Community Survey (ACS) results indicate that during our analysis period, there is a strong relationship between Medicaid enrollment among children and household income. As a result, Medicaid coverage loss among children in our data could be a concern due to two primary factors: increases in household income that result in ineligibility among parents leading to private insurance take-up among their children, and the aforementioned recertification process.[25]

We implement two key sample restrictions to minimize Medicaid coverage gaps while keeping the sample size large enough for our main analysis. First, we restrict to students enrolled in Medicaid in at least half of the years we observe them, which removes the vast majority of coverage gaps (i.e., most of the remaining student record observations match to the Medicaid data). 66% of the remaining sample is always enrolled in Medicaid, with over 80% enrolled over 80% of the

[23] Recall that the 2007 standardized exam was administered in Oct. 2007, so we have no claims leading up to this exam. For the following exam in Oct. 2008, we use all claims starting in Jan. 2008.

[24] These students may be privately insured or they may be uninsured; they may be healthy or they may be unhealthy with private insurance claims that we do not observe.

[25] To deal with these issues, we cannot just use the observations in which we explicitly observe Medicaid enrollment, because then there will be selection into and out of the sample due to transitions into and out of Medicaid coverage. It is also not preferable to restrict only to students that are always enrolled in Medicaid, since some of these students are never eligible for subsidized lunch; with our ultimate goal of identifying heterogeneity in school value-added based on student health, we run the risk of picking up an income effect with this sample. We could restrict to students always enrolled in Medicaid *and* always eligible for subsidized lunch, which would fully eliminate gaps in Medicaid coverage while also removing concern of income effects entirely; however, this would result in a much smaller sample than is preferable (less than 200,000 students) while also inducing more selection of schools into the value-added estimation (schools with few Medicaid-enrolled subsidized lunch-eligible students would be dropped from the analysis). We thus need to implement a combination of strategies to balance Medicaid coverage gaps with sample size.

time. Some of these students, however, are from higher-income households due to Wisconsin's BadgerCare Plus program; to remove concern of an income effect in our main analysis, we next restrict to students eligible for subsidized lunch in at least half of the years we observe them.[26] This also has the added benefit of further reducing gaps in Medicaid coverage: 70% of the final sample is always enrolled in Medicaid, with around 90% enrolled over 80% of the time. With the vast majority of our sample almost always enrolled in Medicaid and from lower-income households, the few Medicaid coverage gaps we do observe are less likely to be from changes in household income, and more likely to be related to automatic dis-enrollment during recertification. We assume no change in health during these lapses in Medicaid coverage. For more details on our sample restrictions and the issues they solve, see Appendix 1.A.

Descriptive Statistics

Table 1.1 presents some descriptive statistics for the sample at each stage of the restriction process with Column (3) representing our final analysis sample for the value-added model. As we move from Column (1) to Column (2) to Column (3), the sample becomes more racially and ethnically diverse with fewer White students and more Black and Hispanic students. Interestingly, our restrictions yield a sample that is actually closer in racial and ethnic composition to 2018 national averages (taken from the American Community Survey and the National Center for Education Statistics) than statewide averages.[27] Increasing the proportion of Medicaid-enrolled and subsidized lunch-eligible students in our sample has the main effect of decreasing average household income; this in turn drives English proficiency down and special education enrollment up from statewide averages. The limited English proficient group is still substantially smaller than the national average, while special education enrollment is much higher. This could be due to transitions into or out of special education for students on the margin; our calculation (which pools all years 2009-2018 together) counts any student with at least one year of special education enrollment in the statistic; the rate of special education enrollment in any single year will be lower.

Another expected effect of our two main sample restrictions is a substantial drop in average test scores: students from lower-income households score around 0.5 SD lower than the statewide mean, which indicates that they score around 1 SD lower than students from higher-income households. This income-based achievement gap has been studied extensively in the economic literature (e.g., Hanushek et al., 2022). While much of the mass in the distribution of our analysis sample is shifted to the left of the population distribution, we still have near-full support, as indicated by the distribution ranges. This means that we still have students from every part of the test score distribution in our sample.

[26] Dynarski and Michelmore (2016) show that students eligible for subsidized lunch consistently are far more likely to be from households that are actually poor than students eligible only once or twice.

[27] While tangential to the main results of this paper, this finding lends credibility to external validity.

Table 1.1: Descriptive Statistics and Sample Restrictions, 2009-2018

	All DPI Students (1)	Medicaid-Enrolled ⩾50% Time (2)	(2) and Elig. for Sub. Lunch ⩾50% Time (3)	2018 U.S. National Averages (4)
Demographic characteristics				
White	0.717	0.527	0.485	0.501
Black	0.105	0.210	0.232	0.143
Hispanic	0.106	0.168	0.183	0.253
Male	0.512	0.510	0.510	0.487
Eligible for subsidized lunch	0.478	0.902	0.981	0.523
Limited English proficient	0.073	0.118	0.130	0.227
Enrolled in special education	0.156	0.241	0.242	0.141
Test scores				
Average grade 5 ELA score	0	−0.419	−0.473	-
[Min, Max] grade 5 ELA score	[-5.24, 7.05]	[-5.24, 7.05]	[-5.24, 7.05]	-
Average grade 8 ELA score	0	−0.433	−0.485	-
[Min, Max] grade 8 ELA score	[-4.51, 5.80]	[-4.51, 5.68]	[-4.51, 5.65]	-
Average grade 11 ELA score	0	−0.502	−0.564	-
[Min, Max] grade 11 ELA score	[-3.10, 3.20]	[-3.10, 2.99]	[-3.10, 2.93]	-
Number of unique schools				
Elementary	855	842	829	-
Middle	579	568	559	-
High	643	621	617	-
Number of unique students	1,176,436	470,408	407,535	-
Number of student-year observations	4,907,079	1,935,581	1,688,164	-

Notes: Column (1) presents statistics for all K-12 public school students in Wisconsin 2009-2018 for whom we observe at least one test score; Column (2) presents statistics for the subset of students enrolled in Medicaid in at least half of the years we observe them; Column (3) presents statistics for our main analysis sample - students enrolled in Medicaid in at least half of the years and eligible for subsidized lunch in at least half of the years we observe them; Column (4) contains national averages. Some of the national averages were taken from the 2018 American Community Survey, other averages were taken from the National Center for Education Statistics for 2018, the final year of our analysis period. For rows containing eligible for subsidized lunch, limited English proficient, and enrolled in special education, we count someone if they ever fit the criteria since there is switching into and out of these groups. Grade 5, grade 8, and grade 11 scores were chosen because these are the primary outcomes of the value-added model. We present ELA test scores for simplicity, but find similar patterns among Math test scores. The rows containing numbers of schools are subject to the constraint that they must have at least 10 students to be counted, the same constraint used in the value-added model.

Lastly, we find that restricting the sample on Medicaid enrollment and subsidized lunch eligibility does decrease the number of schools we are able to look at, though not by much. We subject the school number calculations in Table 1.1 to one of the main constraints of the value-added model: a school is not included if it contains fewer than 10 students.[28] The slight decrease when moving from Column (1) to Column (2) to Column (3) is unsurprising, given longstanding patterns of sorting into schools based on parental income. Interestingly, the pattern is most pronounced in high schools (we lose about 4% of schools), where average school size is actually larger in general. With only 3-4% of each type of school dropping out, we are not worried about having a non-representative sample of schools, even when our sample of students is not representative of Wisconsin.

1.3 Motivating the Use of Health in a Model of Value-Added

Standard Model of School Value-Added

The immediate goal of any model of school value-added is to estimate a credible measure of school quality to be able to make comparisons across schools or across time. Researchers then either work to identify the impacts of changes in quality on longer-term outcomes (e.g., Chetty et al., 2014b, in the case of teacher value-added) or attempt to uncover what sorts of inputs have lasting effects on quality (e.g., Ahn et al., 2023). The latter route is less common in the literature, in part because it is more difficult to pin down exactly what makes schools and teachers more or less effective. In this section we start with a bare-bones model of school value-added, which primarily serves to place schools in Wisconsin - which have not been studied in the value-added literature - among the schools in other states that have been investigated in past work. Next, we begin to introduce student health into the model to demonstrate its relevance to and importance in education research, particularly in the value-added literature. In the following sections we adapt the standard model to allow for heterogeneity in school value-added based on student health, and then identify what school practices impact subgroup-specific school effectiveness.

Though there are a number of different techniques to estimate school value-added, most boil down to a decomposition of students' standardized test scores into components attributed to the school and to student heterogeneity (Rothstein, 2010). Following the recent literature on school effectiveness (e.g., Abdulkadiroğlu et al., 2020), we start with a very simple fixed effect model:

$$Y_{it} = \theta_{j(i,t)} + X'_{it}\beta_X + \omega_t + \varepsilon_{it} \tag{1.1}$$

The outcome is standardized test score - Math and English Language Arts (ELA) - taken in the final year of school enrollment.[29] Due to changing exams and academic standards, we normalize the test score distribution to be mean zero and standard deviation one by grade level and academic year. $\theta_{j(i,t)}$ is a school fixed effect and represents the distribution of key parameters we aim to recover;[30] because it is the case that students attend different schools at different times, we further

[28] In our main analysis (Section 1.5) we pool all years together, so schools are included if there are at least 10 student-year observations associated with that school.

[29] More specifically, the outcome is grade 5 scores for elementary school students, grade 8 scores for middle school students, and grade 11 scores for high school students (students are examined in grades 3-11).

[30] We disallow time-variation to be consistent with subsequent models that will be described.

index schools by i and t to reflect the data generating process (i.e., student i attended school $j(i,t)$ in year t). The comprehensive baseline covariate vector X_{it} is similar to what has been used in Kane et al. (2008) and Chetty et al. (2014a).[31] We control for past test scores using a fully interacted cubic polynomial in past ELA and Math test scores.[32] In addition, we include subsidized lunch status, cumulative years eligible for subsidized lunch,[33] race/ethnicity, gender, lagged absence rate, and indicators for special education enrollment, limited English proficiency, and grade repetition. ε_{it} contains any unobserved student heterogeneity that influences standardized test scores. We estimate the model separately for elementary, middle, and high schools, so we could write Equation (1.1) with all parameters varying by school type; we omit these indices for ease of notation.

Following from past literature, we include separate specifications for ELA and Math test scores as the outcome, and will present results from each specification throughout the paper. Though a subject like Math is highly tracked in high schools (e.g., Figlio and Page, 2002; Hanushek and Wößmann, 2006), this does not interfere with the value-added estimation or interpretation. First, selection of students into advanced Math courses based on past ability will be accounted for in the past test scores that are included in X_{it}. And second, variable selection of students into advanced Math courses driven by differences in schools' course offerings should actually be included in the value-added estimates rather than disentangled from them; these sorts of differences across schools that might drive student achievement are intuitively part of what we think of as school effectiveness. For these reasons, there should be no concern with using Math as the value-added outcome, even though it is a highly tracked subject.

Under the selection on observables assumption, we recover unbiased estimates of $\theta_{j(i,t)}$ and β_X with an OLS regression. This assumption - which will be discussed at length in Section 1.5 - states that school enrollment is as good as random, conditional on observables. We will show that estimates from this baseline model are not perfectly unbiased, due to additional selection of

[31]Though these papers study teacher effects, the identifying assumption of selection on observables is the same. While Abdulkadiroğlu et al. (2020) focus on school effects, they rely more on rank-ordered control functions, so their covariate vector in the OLS estimation is less comprehensive; we therefore opt to include the full set of covariates that have been used in previous work, since the main assumption of our model relies on them. Selection on observables and identification will be discussed in Section 1.5.

[32]Though Chetty et al. (2014a) do not interact past ELA and math test scores, it will help in controlling for heterogeneity across students with strengths in different test subjects. While prior-year scores are appropriate in models of teacher value-added, they will create attenuation bias in the school effect estimates because part of the school's effect may present in earlier grade levels. For this reason, our approach with past test scores is closer to that of Abdulkadiroğlu et al. (2020). We use the most recent scores from the previous school type attended (grade 8 for the high school students and grade 5 for middle school students). For estimating elementary school value-added, using grade 3 scores could create attenuation bias if part of the impact of elementary schools on student achievement occurs prior to grade 3; to solve this issue we use grade 3 scores divided by two which will be somewhere between the student endowment and the level of human capital at the end of grade 3. The idea is to approximate student ability at the start elementary school, otherwise comparisons across school types are impossible; in any case, the elementary school value-added results are robust to using grade 3 scores, grade 3 scores divided by two, and zero exactly (which assumes no human capital accumulation prior to test-taking years). Future work could explicitly model human capital accumulation to predict what test scores would be at the start of elementary school.

[33]Though Chetty et al. (2014a) do not include cumulative years eligible for subsidized lunch, Dynarski and Michelmore (2016) find there to be important heterogeneity in test scores within the lower-income group, based on duration of eligibility for subsidized lunch. They also find that students consistently eligible for subsidized lunch are far more likely to be from low-income households than students eligible once or twice.

students into schools based on health.

Controlling for Student Health in the Value-Added Model

We will adapt the model in Equation (1.1) in two ways. First, we control for health as an additional dimension of student heterogeneity. The motivation for this extension stems from ongoing debate regarding the credibility of the selection-on-observables assumption that underlies models of value-added: researchers often ask whether X_{it} is enough to control for student heterogeneity and identify the school fixed effect parameters. And second, we will construct subgroups to distinguish between healthy and unhealthy students, which will be interacted with the school fixed effect to yield measures of school effectiveness that vary based on student health; we will carry out this extension in Section 1.5. In Equation (1.2), we supplement the demographic covariate vector X_{it} (identical to the vector in Equation (1.1)) with a health covariate vector H_{it} to control for additional student heterogeneity along a dimension that is typically unobserved in education data:

$$Y_{it} = \theta_{j(i,t)} + X'_{it}\beta_X + H'_{it}\beta_H + \omega_t + \varepsilon_{it} \tag{1.2}$$

In H_{it} we include students' health-related absence rates that are predicted in a random forest model (described in the following section), as well as lagged and twice-lagged predicted health-related absence rates. These predictions give us a way to proxy current health as well as any changes in health that may have occurred since the realization of the past test score that we control for in X_{it}. We also include the observed absence rate, so that we can disentangle the main effect of poor health (which comes from predicted absence rates) from the main effect of absence more broadly. In addition, H_{it} contains number of emergency room (ER) visits and Medicaid claims, and indicators for diagnosis codes from ER, inpatient, and outpatient claims.[34] As in the previous specification, test scores are the dependent variable.

Even if the combination of demographic characteristics controlled for in the standard model in some way proxies health, it is likely that there is additional sorting of students into schools based on health (i.e., there are more unhealthy students in lower value-added schools), which would suggest inflation in magnitude of school effects that have been conflated with selection of students into schools based on health. In this regard, we may find that past work has upwardly biased estimates of school effect dispersion, and it would be important to control for student health when possible. It is also unclear whether the ranking of schools is preserved across the specifications. The concern is the following: schools that are estimated be the highest value-added in the standard model may contain the healthiest students, so once we control for student health, are these schools still estimated to be the most effective at generating test score gains?

[34]For our sample, the Medicaid claims data contain over forty thousand unique diagnosis codes; controlling for all of these (plus lags) poses some serious technical issues. Instead, we take only the diagnosis codes determined to be above mean importance in the random forest; these are the codes that most effectively split the sample in terms of absence rate (many - though not all - of these are also likely to be the most relevant for test scores too). This gets the covariate set in H_{it} down to only a couple thousand codes.

Measuring Student Health: Random Forest

We now describe how we measure student health for the purpose of implementing the first extension to the model that was discussed in the previous section.[35] As mentioned in Section 1.2, our Medicaid claims data contain ICD-9 and -10 diagnosis codes, which are "used by physicians to classify and code all diagnoses, symptoms, and procedures for claims processing," according to the American Medical Association. We quantify student health from these data with a random forest, a supervised machine learning technique that allows us to predict absence rates in school directly from observed ICD diagnosis codes and student age.[36] We denote the predicted absence rates as a health index because they have the simple interpretation that a higher value signifies higher risk of health-related absence from school. Currently, this represents a univariate measure of student health, though our method can be adapted with ease to a multivariate setting by combining multiple outcomes together in the random forest.

There are several important reasons we need to quantify health in the first place, and why we choose to implement a random forest specifically. With over forty thousand unique diagnosis codes spanning over twenty five million Medicaid claims, it is simply econometrically infeasible to include all of the diagnosis codes directly in the value-added model. As a result, we need a reliable way to transform an immense amount of data into a more accessible measure of health, which we will later use to distinguish between healthy and unhealthy students. With a continuous outcome like absence rate, we face a prediction problem: what absence rate would we predict for a student with diagnosis codes x, y and z? Machine learning techniques are particularly well-suited to prediction problems, because they allow for some bias in-sample to generalize better out-of-sample (Kleinberg et al., 2015; Obermeyer and Emanuel, 2016; Mullainathan and Spiess, 2017).[37] The random forest does notably well to handle large numbers of sparse predictors, combining them in nonlinear and interactive ways, allowing us to assign students with similar diagnosis codes similar health indexes and ultimately distill tens of thousands of diagnosis codes into a more intuitive measure.

For the remainder of this section we outline the random forest model briefly; more details on the application of the random forest to our context can be found in Appendix 1.B, and further information can be found in Breiman (2001) and Biau and Scornet (2016). At the highest level, a random forest is a collection of B decision trees. Each decision tree b takes a random subset of diagnosis codes (predictors) to recursively split a random sample S_b of student-year (i, t) observations. This is done until tree depth d is reached. For ease of notation, we let $k = (i, t)$ denote the student-year tuple. At this point, the observation k ends up in leaf node $L_b(k)$ based on its predictor set X_{it}^{RF}, and the decision tree returns a prediction equal to the average absence rate among all observations in subsample S_b that end up in the same leaf node as k. The random forest then averages over decision tree predictions to make the final prediction $\hat{\mu}^{RF}(k)$ for each observation k.

[35]Section 1.4 will outline how we use our measure of student health to distinguish between healthy and unhealthy students. In Section 1.5, we will carry out the second model extension.

[36]Robustness to absence rates as the random forest outcome are contained in Appendix 1.B.

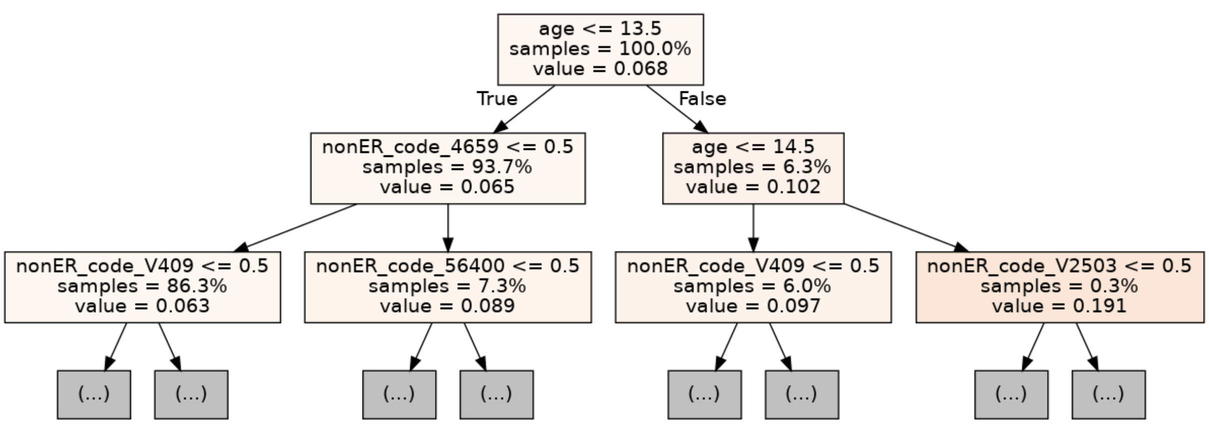
[37]High performance out-of-sample makes machine learning techniques generalizable to a number of different contexts and populations, should that be where researchers have interest.

Borrowing notation from Athey and Wager (2019), we denote the random forest's prediction for observation k (representing student i in year t) more formally by the following average:

$$\hat{\mu}^{RF}(k) = \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in S_b}\frac{Y_{it}^{RF}\mathbb{1}\{k \in S_b, X_{it}^{RF} \in L_b(k)\}}{|\{k : k \in S_b, X_{it}^{RF} \in L_b(k)\}|} \tag{1.3}$$

The inner sum applies to individual decision trees, averaging over absence rates of observations that end up in the same leaf node based on their set of diagnosis codes. The outer sum averages over the B decision trees' predictions to give the random forest prediction $\hat{\mu}^{RF}(k)$ for student i in year t. In our context, we estimate an absence rate prediction directly from current academic year diagnosis codes and previous academic year diagnosis codes; we also include student age and construct separate models for elementary, middle, and high school students because the health production function is not constant over time. We exclude other observables like demographic characteristics from the model to force the random forest to pick up variation primarily driven by differences in health. For this reason, we refer to the predictions yielded from the random forest as the predicted health-related absence rate or simply as our constructed health index. Because we include lagged diagnosis codes in the random forest, we are unable to make predictions for students in 2008; we recover health-related absence rate predictions for all students from 2009 through 2018, so this is our main analysis period for the estimation of school value-added.[38]

Figure 1.1: Decision Tree Snippet



Notes: Each node represents a decision made by the tree and contains the model predictor used in splitting the sample, the threshold to maximally split the sample (for indicator variables like diagnosis codes, the threshold is 0.5 by construction), the percent of the sample to which the current path through the tree applies (in "samples"), and the current prediction if an observation were to stop at that node (in "value"). This tree only has the top three levels displayed, but it grows far deeper.

[38] Though we are restricted to Medicaid enrollment and claims, we are able to recover predictions for students when they are not covered by Medicaid because we assume no change in health during gaps in Medicaid coverage (diagnosis codes are the same as in the last period of Medicaid coverage). With our set of sample restrictions described in Section 1.2, we observe very few gaps in Medicaid coverage for our sample.

To make the theory behind a decision tree clearer, we now walk through a brief example for the reader. In Figure 1.1, we present the top three levels of a single decision tree used in the random forest for middle school students; we denote this tree by B_1. Each observation in the sample that is randomly chosen by B_1 is fed into the tree, starting in the uppermost node. A random set of predictors are chosen, with student age representing the first one; the value 13.5 is chosen to maximally split the sample. Observations that satisfy the splitting criterion (age $\leqslant$ 13.5) are sent into the left half of the tree (93.7% of the observations randomly selected by B_1 satisfy the criterion) while observations that do not satisfy the criterion are sent into the right half; the procedure is repeated until the maximum depth of the tree is reached. At each level, the predicted absence rate for each observation is updated and contained in the "value" of each node. B different decision trees that select different samples and predictor sets are constructed, and the random forest averages over these predictions to predict a final absence rate for each student in each year.

Table 1.2 reports the means and standard deviations of our constructed health index separately by school type as well as means and standard deviations for absence rates that are observed in the data. The random forest model predicts a distribution very close in mean to what is observed, which is expected because it relies on repeated averages over observed absence rates for various subsamples. The variance, however, is not well-matched; variation in the health index is about a quarter to a third of the variation in observed absence rate. We attribute this discrepancy to the fact that there are many other factors aside from diagnosis codes (health) that are predictive of absences, such as demographics and parent health; Section 1.5 outlines why we are not concerned about this. The idea that we are primarily picking up variation in absence rates that is generated by student health (represented by the diagnosis codes) is why we refer to our constructed index as the predicted health-related absence rate. More patterns in the health index along with additional descriptive statistics are contained in Appendix 1.B.

Table 1.2: Descriptive Statistics on Student Health

	Elementary school		Middle school		High school	
	Mean (1)	SD (2)	Mean (3)	SD (4)	Mean (5)	SD (6)
Student-level health index	0.056	0.013	0.078	0.031	0.117	0.049
Observed absence rate	0.056	0.057	0.078	0.088	0.119	0.151
School-average health index	0.056	0.004	0.077	0.008	0.114	0.023
School-average absence rate	0.053	0.019	0.072	0.031	0.101	0.074

Notes: This table reports means and standard deviations (SD) in our constructed health index compared to observed absence rates in the data. We also average the student health measure and absence rates by school, and produce means and standard deviations of the school-level distributions.

Evidence of the Importance of Health

In this section we present some of the results derived from the two models of school value-added discussed thus far. Table 1.3 shows how the dispersion (standard deviation) in school effects changes when we adjust the specification to include student health.[39] In elementary schools, we find that dispersion is biased upward by 2%; in middle schools, we find dispersion to be biased upward by around 4%; in high schools, dispersion is biased upward by as much as 5%. While these differences are not very large, they are also nontrivial. Bias in high school dispersion is statistically different from zero, indicating that models that omit health (e.g. Angrist et al., 2017; Abdulkadiroğlu et al., 2020) recover estimates that are biased upward in magnitude, which overstates the role of the school in generating test score gains. Furthermore, because we have heavily restricted the student sample to get the best match to our Medicaid claims and enrollment data,[40] we have removed some student heterogeneity from the models that creates additional variation in the school effects. As shown in Appendix 1.C, using the population rather than the lower-income sample yields very similar estimates on dispersion when student health is omitted, meaning that selection of students into our sample is not a concern for the baseline model. However, there is more health heterogeneity in the population, which means that the estimated difference in dispersion *across* models is likely a lower bound and would be larger if one were to include private insurance claims.

Table 1.3: Dispersion (Standard Deviations) in School Effects Across Models

	Elementary school		Middle school		High school	
	Standard Model (1)	Health Included (2)	Standard Model (3)	Health Included (4)	Standard Model (5)	Health Included (6)
	Outcome: ELA test scores					
Dispersion	0.119 (0.003)	0.117 (0.003)	0.170 (0.005)	0.164 (0.005)	0.149 (0.004)	0.144 (0.004)
Number of schools	828	828	558	558	606	606
	Outcome: Math test scores					
Dispersion	0.156 (0.004)	0.154 (0.004)	0.184 (0.006)	0.179 (0.005)	0.138 (0.004)	0.132 (0.004)
Number of schools	828	828	558	558	606	606

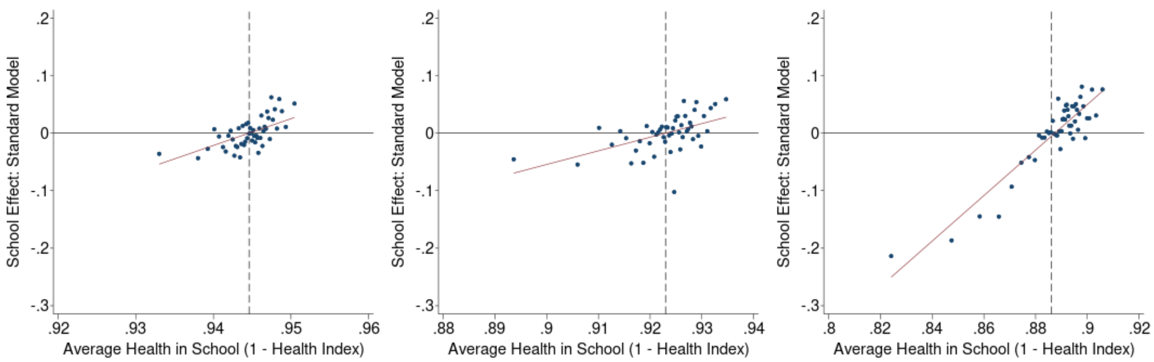
Notes: This table reports dispersion of school effect estimates from two models of value-added. Estimates in Columns (1), (3), and (5) come from the standard model in Equation (1.1) while those in Columns (2), (4), and (6) come from the model that controls for health, described in Equation (1.2). Standard errors are calculated following Ahn and Fessler (2003), though one could also bootstrap the standard errors.

[39]Table 1C.2 in Appendix 1.C compares coefficient estimates from some of the key covariates in the X_{it} and H_{it} vectors.

[40]We do this even in the standard model. This way, the *only* change between the two models is that we control for health in one of them. Changing the sample and the controls across specifications would prevent us from identifying the true effect that controlling for health has on the model estimates.

The finding that dispersion in value-added models that leave health out are biased upward implies that the magnitudes of the school effects are also biased upward. This suggests a systematic relationship between value-added in the standard model and student health. We show in Figure 1.2 that there is a positive correlation between the standard estimate of school effectiveness and school-average student health.[41] This pattern is most notable in high schools, where we find the largest discrepancy between estimates from the two models. This systematic relationship indicates selection of students into schools based on health (i.e., healthier students select into higher value-added schools), an intriguing pattern that future work should further explore to more fully understand.

Figure 1.2: Correlation in Standard School Effects with Student Health



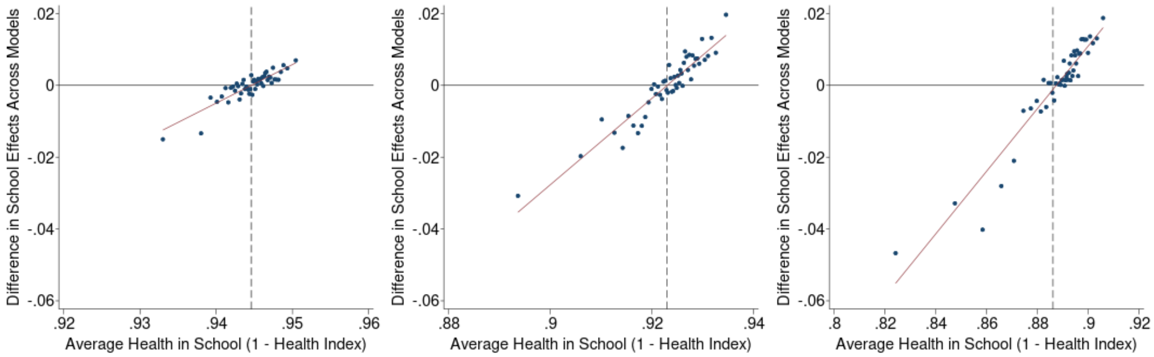
(a) Elementary (b) Middle (c) High

Notes: On the x-axis is school-average health (one minus the health index to give the interpretation that larger numbers mean healthier students), and on the y-axis is the estimated school effect in the standard model that leaves health out. The vertical dashed line is the mean school-level health average across schools. These are binscatters with 100 groups, to give a sense of the broader pattern.

Under the selection on observables assumption, selection bias in the value-added estimates decreases by increasing the set of controls in the model, so estimates from the model that controls for student health are less biased and closer to the truth. We supplement Figure 1.2 with an additional set of graphs in Figure 1.3, which plot the difference in value-added estimates across models. This difference measures bias in the estimates from the standard model. A negative difference indicates that the standard model underestimates true school effectiveness, while a positive difference indicates that the standard model overestimates true effectiveness. We find a fairly strong positive relationship between the bias and student health that grows from elementary to middle to high school. This means that schools with worse average student health have effectiveness that is underestimated in the standard value-added model; schools with better average student health have effectiveness that is overstated. The findings from this section indicate (1) that there is clear sorting of students into schools based on underlying health, and (2) that systematic bias in school effect estimates arises due to the conflation of school effectiveness with heterogeneity in student health, which leads to inflated estimates of dispersion in school value-added.

[41]We calculate "health" to be one minus our constructed health index, which gives the more intuitive interpretation that a larger number signifies better health.

Figure 1.3: Explaining the Difference in School Effects Across Models with Student Health



(a) Elementary (b) Middle (c) High

Notes: On the x-axis is school-average health (one minus the health index to give the interpretation that larger numbers mean healthier students), and on the y-axis is the difference in school effects across models. This is constructed to be the estimate from the standard model minus the estimate from the model that includes health. The vertical dashed line is the mean school-level health average across schools. These are binscatters with 100 groups, to give a sense of the broader pattern.

Table 1.4: Correlation in School Effect Estimates and the Health Index

	School effects: standard model (1)	School effects: health included (2)	School-average health (3)
		Elementary school	
School effects: standard model	1.000		
School effects: model with health	0.995	1.000	
School-average health	0.064	0.049	1.000
		Middle school	
School effects: standard model	1.000		
School effects: model with health	0.990	1.000	
School-average health	0.077	0.038	1.000
		High school	
School effects: standard model	1.000		
School effects: model with health	0.984	1.000	
School-average health	0.244	0.192	1.000

Notes: This table reports correlations of school effect estimates across value-added models (with and without controlling for health) and correlations of school effects with student health. In each panel, the [2,1] entry is the correlation across the two models; the [3,1] entry represents the correlation between the standard school effect and student health; the [3,2] entry represents the correlation between the other school effect and student health. For the purpose of recovering these correlation coefficients, the value-added estimates have been normalized to be mean zero and standard deviation one by school type.

To conclude this section, in Table 1.4 we provide correlation coefficients between the school effects estimated in the standard model, the school effects estimated in the model that accounts for health, and our measure of school-average student health. There is a strong positive correlation between estimates from the two models. The school effects from the standard model are positively correlated with student health; this correlation is weak in elementary and middle schools and slightly stronger in high schools. These positive correlations signify selection on health; they are what ultimately drive upward systematic bias in the magnitude of school effects in the standard model. Once we control for student health in the model, the correlation between the estimated school effects and student health diminishes, indicating that some of the student heterogeneity in health is wrongly attributed to the school effects in the standard model. This section has uncovered evidence of the importance of controlling for student health in models of school value-added; we now show that there is also important heterogeneity in school effectiveness based on student health.

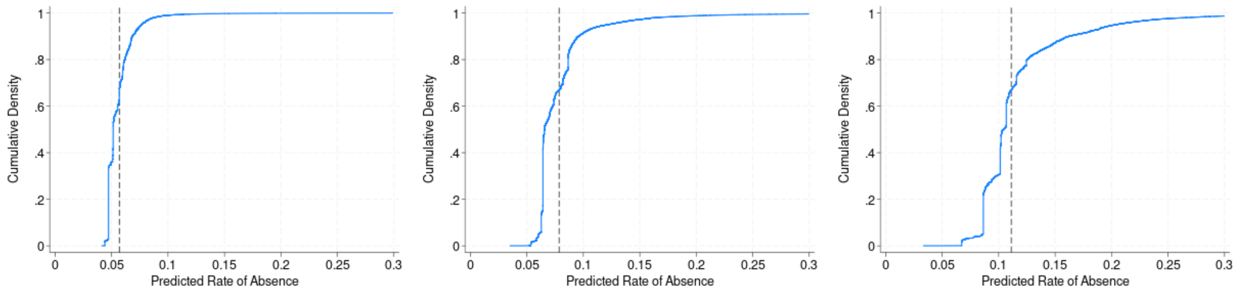
1.4 Constructing Health-Based Subgroups

For the purpose of identifying heterogeneity in school effectiveness based on student health, we form health-based subgroups from our constructed health index. Our primary goal is to classify each student as either healthy or unhealthy; to do this, we discretize the predicted health-related absence rates to form two groups of students. We need the threshold between groups to be high enough that students that are actually healthy do not end up in the unhealthy group; in other words, we want the unhealthy group to be composed of students that are *actually* likely to be unhealthy. We also need the threshold between groups to be low enough that the unhealthy student group has a large enough sample size to mitigate sampling error in the value-added estimation that stems from small cell sizes. Since the predicted absence rate distributions are heavily skewed right with substantial mass near the mean, we use a two-thirds/one-third split to place most students in the healthy group. With this threshold, we are likely to target students that are actually unhealthy in the unhealthy group.[42] Since we estimate value-added separately by school type - and with the health production function likely varying across students of varying age - we construct health groups separately by school type (elementary, middle, and high). To allow health to vary from year to year, we identify the health group thresholds using all years of health indexes pooled together.[43] In Figure 1.4 we show cumulative density functions of the absence rate predictions, separated by the health group threshold. For elementary school students the threshold is estimated to lie at a predicted absence rate of 0.057; for middle school students the threshold is at a rate of 0.079; for high school students the threshold is at a rate of 0.111. Increases from elementary to middle to high school are partly due to higher absences among older students, partly due to more severe health conditions, and partly due to the way health maps into absences.

[42] Other nearby thresholds yield similar results for our main analyses.

[43] Consider, for example, the case in which everyone is less healthy than they were in the previous year. The unhealthy group should grow across these two years to capture changes in health over time.

Figure 1.4: CDFs of the Health Index by School Type



(a) Elementary (b) Middle (c) High

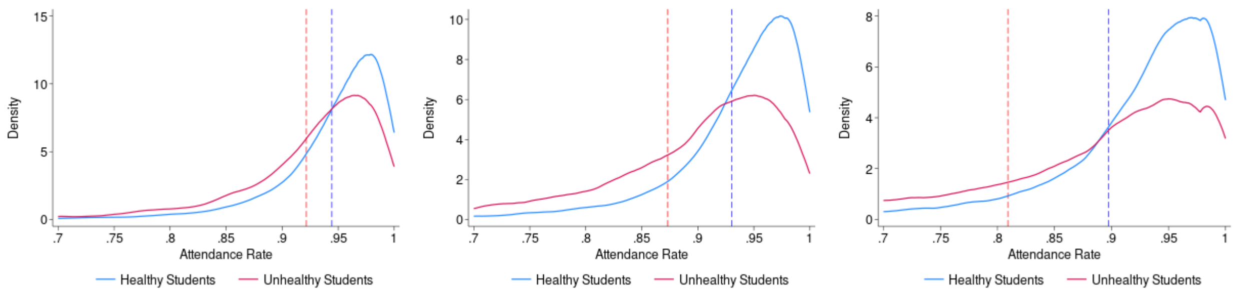
Notes: These graphs presents cumulative density functions of random forest-predicted absence rates separately by school type. The dashed vertical line splits students into health groups, with roughly two-thirds in the healthy group and one-third in the unhealthy group.

There are a number of reasons to use categorical health groups instead of the continuous health index in the value-added model. The first is that we rely on the assumption of linearity in health when using the continuous index directly in the model. We expect this not to hold. For students at the lower end of the health distribution (healthy), the marginal effect of poor health likely has little effect on students' test scores; for students at the higher end of the health distribution (unhealthy), the marginal effect of poor health might have more drastic effects on academic outcomes.

The second reason against keeping things continuous is measurement error. One issue with the random forest is that it will predict a low health index for students that never go to the doctor when they are sick, and as a result these students may look healthier than they really are. Another issue is for unhealthy students that miss school but don't go to the doctor; this will push the mean absence rate up among students with no diagnosis codes, affecting the left tail of the predicted absence rate distribution (and thus inflating predictions among students that have no diagnosis codes because they truly are the healthiest). While constructing health groups does not eliminate measurement error, it does give us fewer things to worry about. With the continuous health index, we are concerned with error at *every* part of the predicted absence rate distribution. When we construct health groups, we are now only concerned with measurement error that induces jumps over the health group threshold. First, for students near the margin there is very little we can do to avoid issues associated with measurement error, other than using other outcomes for the random forest to validate the health group determinations from absence rates (e.g., emergency department visits). Second, because our threshold generates such a large healthy group, health indexes for students that are far from the threshold are unlikely to contain enough measurement error that estimated health group placement is incorrect. In other words, because students that don't go to the doctor when they are sick are probably not the healthiest but also probably not the unhealthiest, they will likely be placed in the healthy group regardless; inversely, students that truly are unhealthy are more likely to go to the doctor and receive a diagnosis, and thus will be placed in the unhealthy group with our method.

To verify that we are separating healthy from unhealthy students, the next three figures present distributions of observed academic and health outcomes by health group. Each graph contains students in the 2018 academic year; all observations in this year are in the test set of the random forest, which means that they are not used to "train" the model how to map diagnosis codes to absence rates.[44] In other words, the health groups in this year come from true predictions of the random forest, so differences across groups are not mechanical. Figure 1.5 shows that observed attendance rates are substantially lower among unhealthy students, especially in high schools.

Figure 1.5: Observed Attendance Rate by Health Group (2018)

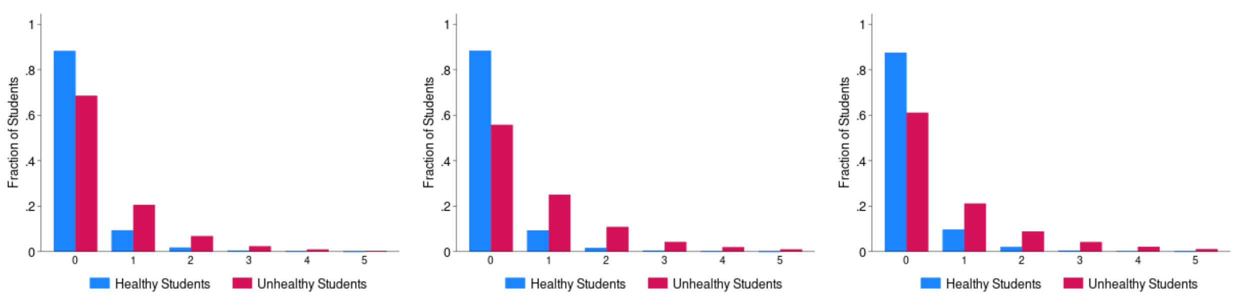


(a) Elementary (b) Middle (c) High

Notes: These graphs present observed attendance rate distributions in 2009 among students in each health group. The blue line contains healthy students while the red line contains unhealthy students. We chose 2009 because it is the first year of the main analysis window; we find similar patterns in other years.

Figure 1.6 shows that differences in predicted absence rates across health groups also carry over to health outcomes; overall, a much higher fraction of students that were placed in the unhealthy group visited the emergency department more often than students that were placed in the healthy group. The fact that we find similar patters in a more objective health outcome is a good verification of our absence rate health measure.

Figure 1.6: Emergency Department Visits by Health Group (2018)



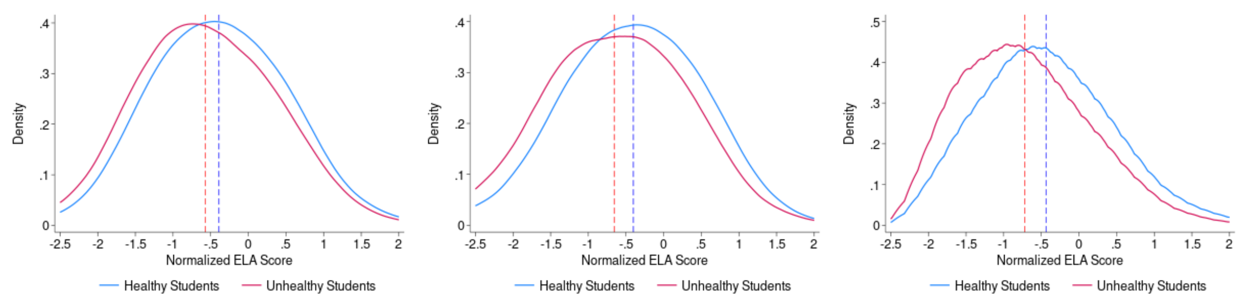
(a) Elementary (b) Middle (c) High

Notes: These graphs present the frequency with which students from different health groups visited the emergency department in 2009. The blue bars contain healthy students while the red bars contain unhealthy students. We find similar patterns in other years and when looking at total Medicaid claims instead.

[44] In Appendix 1.C we also show differences across health groups for 2009, a training set year.

Figure 1.7 shows that unhealthy students score lower on standardized exams than healthy students (we present only ELA test score distributions for brevity, but find the same pattern in the Math test score distributions). At the mean, unhealthy students in our sample score substantially lower than healthy students - on top of the already substantial gap between students from lower-income households in our sample and students from higher-income households that were been dropped during sample construction (shown in Table 1.1). In elementary schools, the health-based achievement gap is smaller (only around 0.2 SD); the gap grows to around 0.25 SD in middle schools and 0.3 SD in high school. This means that the health-based achievement gap among high school students from lower-income households is at least a third of the magnitude of the income-based gap statewide; the health-based gap is actually smallest in 2018, so these are lower bounds. With students that are from lower-income households *and* are unhealthy, the gap compounds.

Figure 1.7: Normalized ELA Test Scores by Health Group (2018)



(a) Elementary (b) Middle (c) High

Notes: These graphs present ELA test score distributions in 2009 among students in each health group. The blue line contains healthy students while the red line contains unhealthy students. We find similar patterns in other years; using Math test scores we find even greater separation between health groups.

In Table 1.5 we present descriptive statistics by health group for years at the beginning and end of our analysis window (switching across health groups from year to year prevents us from pooling years). We find that while the unhealthy group consistently has fewer white students and more black students, the groups are generally more racially and ethnically similar to each other than to the greater Wisconsin student population. Notably, test scores among unhealthy students are considerably lower than among healthy students; because the groups are demographically not so different from each other, we argue that a large part of this apparent achievement gap is truly due to student health. Routine check-ups, vaccinations, myopia, and pharyngitis represent the most common diagnosis codes among students in the healthy group while routine check-ups, asthma, pharyngitis, and upper respiratory infection more broadly represent the most common diagnoses among students in the unhealthy group. It is unsurprising that the most common diagnosis codes are similar across health groups, because the most frequent codes are often the least severe. What differs more across groups is the number of diagnosis codes per claim: healthy high school students have 2.16 codes per claim while those in the unhealthy group have 2.56 codes per claim.

Table 1.5: Descriptive Statistics by Health Group

	2009 Academic Year			2018 Academic Year		
	All students (1)	Healthy students (2)	Unhealthy students (3)	All students (4)	Healthy students (5)	Unhealthy students (6)
Demographics						
White	0.547	0.568	0.500	0.476	0.482	0.469
Black	0.247	0.218	0.314	0.190	0.173	0.218
Hispanic	0.126	0.126	0.128	0.214	0.216	0.209
Male	0.505	0.498	0.520	0.511	0.523	0.488
Elig. for subsidized lunch	0.921	0.913	0.937	0.866	0.843	0.904
Limited English prof.	0.083	0.089	0.071	0.096	0.099	0.087
Enrolled in special ed.	0.233	0.189	0.336	0.198	0.161	0.269
Test scores						
Grade 5 ELA score mean	−0.495	−0.432	−0.653	−0.458	−0.390	−0.569
Grade 8 ELA score mean	−0.481	−0.343	−0.688	−0.464	−0.402	−0.656
Grade 11 ELA score	−0.540	−0.433	−0.929	−0.558	−0.438	−0.720
Number of unique students	52,807	36,732	15,892	68,158	43,588	23,516

Notes: This table presents descriptive statistics by health group for students in our main analysis sample for the value-added estimation; due to the switching across health groups through time (Table 1C.3 in Appendix 1.C shows changes in health group membership from year to year), we cannot pool years like we did in Table 1.1. We present 2009 and 2018 to show what the groups look like at the beginning and end of our data; other years look similar. The achievement gap is decreasing over time.

1.5 Health-Based Heterogeneity in School Effectiveness

Model of School Value-Added

We modify the specification from Equation (1.2) with an interaction between the school fixed effect and a health group fixed effect, which allows for heterogeneity in school value-added based on student health. The idea is to estimate the subgroup-specific value-added of each school to understand whether school j's effectiveness varies across student subgroups determined by health. If there are differences across subgroups, we would have evidence of match effects: some schools are more effective at boosting test scores for certain types of students. The econometric model - which we estimate separately for elementary, middle, and high schools - is written as follows:

$$Y_{it} = \theta_{j(i,t),h(i,t)} + X'_{it}\beta_X + H'_{it}\beta_H + \omega_t + \varepsilon_{it} \tag{1.4}$$

On the left-hand side are students' standardized test scores Y_{it}. The first term on the right-hand side $\theta_{j(i,t),h(i,t)}$ is a school fixed effect $(j(i,t))$ interacted with a health group fixed effect $(h(i,t))$,[45] and encompasses the subgroup-specific impact of schools (i.e., their value-added); these are the distributions of parameters that we are interested in recovering. Health groups $h(i,t) \in \{0,1\}$ are

[45]Though the health groups are constructed objects from predicted values, we write out econometric models for the population. For this reason, it isn't appropriate to include a hat on h, even though it is estimated. As the sample size grows large, our random forest predictions and thus health group estimates approach the true values in the population.

determined by random forest predictions of absence rates, as discussed in Sections 1.3 (outline of the random forest model) and 1.4 (construction of the health-based subgroups); $h = 0$ represents the unhealthy group (students more likely to miss school due to health) while $h = 1$ represents the healthy group.[46] Schools and health groups are further indexed by i and t to reflect two key points: (1) students attend different schools at different times and switch health groups from time to time (i.e., student i in health group $h(i,t)$ at time t attends school $j(i,t)$ at time t), and (2) what matters for school j's value-added are the students in health group h that attend school j at the time at which they have outcome test scores realized.[47] For ease of notation, we drop these additional indices for the remainder of the paper, simply denoting $j(i,t)$ by j and $h(i,t)$ by h.

The covariate vectors X_{it} and H_{it} control for student heterogeneity; X_{it} is identical to what was described in Section 1.3, and H_{it} has been modified slightly to better disentangle subgroup-specific school value-added from the main effect of poor health on test scores. In X_{it}, we control for past test scores using a fully interacted cubic polynomial in past ELA and Math test scores; in addition, we include subsidized lunch status, cumulative years eligible for subsidized lunch, race/ethnicity, gender, lagged absence rate, and indicators for special education enrollment, limited English proficiency, and grade repetition.

In H_{it}, we include a full interaction between the predicted health-related absence rate, resulting health group placement, and past test scores, which helps in removing the main effect of poor health from the school effects and allows for flexibility in health heterogeneity among similar-ability students. We also include lagged and twice-lagged predicted health-related absence rates, which give us a way to proxy for any changes in health that may have occurred since the realization of the past test score controlled for in X_{it}. We include observed absence rate so that the main effect of absence is separated from the subgroup-specific school effects (which are determined from predicted absence rate). Lastly, we include number of ER visits and Medicaid claims, and indicators for important diagnosis codes from ER, inpatient, and outpatient claims, which are more objective health measures that come directly from the raw administrative data.

We include a year fixed effect ω_t to absorb any time trends. The error term ε_{it} contains any variation in test scores that is unexplained by schools or observable student characteristics controlled for in X_{it} or H_{it}. By including a rich set of demographic characteristics, health characteristics, and lagged test scores, we try to remove as much student heterogeneity from the unobservable error

[46] If we interact the continuous health index with the school fixed effect (and also include an un-interacted school fixed effect) instead of using the constructed health groups, we would recover a school effect on students with a zero prediction and school-varying marginal effects of poor health. Though appealing, there are some pitfalls to this method (aside from the larger concern of measurement error discussed in the previous section). This model would rely on the assumption of constant marginal effects across the distribution of student health, which may be unreasonable if the relationship between health and academic outcomes is nonlinear (i.e., the marginal impact of a worsening of health on academic outcomes is very negative for students that are more unhealthy while it may be near zero for students that are quite healthy). In addition, the interaction between the school fixed effect and the continuous health index would inflate the standard errors, which may be an issue during the Empirical Bayes shrinkage procedure.

[47] The outcome is grade 5 test scores for the estimation of elementary school value-added, grade 8 test scores for middle school, and grade 11 test scores for high school (the final year of enrollment at each school type). For elementary school j's value-added on healthy students ($h = 1$), the students that factor into the estimation are those that are healthy and attend school j in grade 5.

term as possible; nevertheless, there will still be some unobserved ability and health. We now show why concern over unobservables is minimal.

We de-mean the school-by-subgroup fixed effect estimates separately by subgroup (and school type) such that each distribution is mean zero; prior to de-meaning, the school effect estimates are centered around the mean subgroup-specific outcome, so this de-meaning is necessary for the value-added interpretation. The effect of the mean school on subgroup h is zero; less effective schools have negative effects while more effective schools have positive effects. We first take the mean separately by health group:

$$\bar{\theta}_h = \frac{1}{J}\sum_j \left[\theta_{jh}\right] \text{ for } h = 0, 1 \tag{1.5}$$

Each fixed effect estimate is then shifted by the relevant mean. E_{jh} represents the effect of school j on test scores of students in health group h:

$$E_{jh} = \theta_{jh} - \bar{\theta}_h \text{ for } j = 1, ..., J \text{ and } h = 0, 1 \tag{1.6}$$

Identification: Selection on Observables

The primary assumption underlying any model of school value-added is selection on observables, under which an ordinary least squares (OLS) regression yields unbiased estimates of θ_{jh}, β_X, and β_H.[48] In our context, we require that the combination of school enrollment and health group placement are as good as random, conditional on the covariate vectors X_{it} and H_{it}. Formally, we write the assumption as follows:

$$E[\varepsilon_{it} \mid j, t, X_{it}, H_{it}] = 0, \quad \text{for } j = 1, ..., J, \text{ and } h = 0, 1 \tag{1.7}$$

In words, all sorting of students into schools and health groups is observed and controlled for in the model. Most important in controlling for sorting into schools are past test scores and cumulative subsidized lunch eligibility. Past test scores are crucial in controlling for underlying student ability as well as any unobservable inputs to student ability (for example, parental resources). Cumulative subsidized lunch eligibility helps in identifying longer-term household income that in large part determines residential location (which determines enrollment in public schools in Wisconsin). Though there has been long-standing debate over the credibility of the selection-on-observables assumption in models of school value-added (Rothstein, 2010, 2017; Chetty et al., 2014a,b), recent work leveraging lotteries to assess bias has found that while conventional value-added estimates are biased, decisions on policy informed by these models are likely to boost student achievement (Deming, 2014; Angrist et al., 2016, 2017). Furthermore, Chetty et al. (2014a) and Angrist et al. (2023) find that bias of the value-added estimates is limited when controlling for students' past test

[48]In our case, we are only interested in recovering θ_{jh}, which means selection-on-observables is sufficient but not necessary. All we need is orthogonality between θ_{jh} (residualized by X_{it} and H_{it}) and ε_{it}.

scores; Abdulkadiroğlu et al. (2020) also find low bias when comparing to a rank-ordered control function method, even when using a much more limited set of controls in the X_{it} vector.

In dealing with sorting into health groups, we crucially must control comprehensively for student health. Current health is primarily picked up by the health index (predicted absence rate) and diagnosis codes from ER, inpatient, and outpatient claims; changes in health between outcome test scores are picked up by lags of the health index. In this regard, this part of selection on observables boils down to the assumption that the health index lags adequately proxy for health in previous years so that any changes in health are properly controlled for. Demographic characteristics - which likely also affect student health but are not included in the random forest - are taken care of in the X_{it} vector. One final concern is that there may be additional factors that affect students' health-related absences (and hence health group placement) that are not controlled for in the value-added model - nor in the random forest and are therefore not encompassed by the health index - such as parent health. Though parent health likely affects student absences on the lower end of the distribution, it is common for chronically ill parents to make alternative arrangements to prevent a pattern of chronic absence that is a direct result of poor parental health. Furthermore, in our analysis, we place the majority of students in the healthy group, so parent health likely has no effect on the child's health group placement even if it influences their estimated health index.

In one test to assess the degree of selection on unobserved factors, Chetty et al. (2014a) bring into the value-added model additional controls not included in the baseline specification. When moving from the basic model without health in Equation (1.1) to the more comprehensive model with health in Equation (1.2), we find a strong correlation between estimates across models.[49] This would indicate that even though there is selection into schools based on health, most of the sorting is picked up by demographic characteristics and the bias is not substantial. In our case, since we control for health with the rich H_{it} vector, we are even less concerned about unobserved factors violating the selection-on-observables assumption regarding health group placement.[50]

Dispersion Results

In this section we present results from the main specification of our model of school value-added, which is described in Equation (1.4). In particular, in this section we explore dispersion in school effectiveness as well as correlations across health subgroups. In Appendix 1.C, we summarize coefficient estimates for several of the key covariates in the X_{it} and H_{it} vectors that control for student heterogeneity and allow for identification of the school effects. Table 1.6 summarizes our estimates of dispersion in schools effectiveness on students of varying underlying health (these are estimates of the standard deviation of value-added across schools).[51] With the exception of middle

[49]The correlation in school effect estimates across the two models is 0.995 for elementary schools, 0.990 for middle schools, and 0.984 for high schools.

[50]For more thorough testing of the assumption, one could go the route of Altonji et al. (2005, 2008).

[51]In Appendix 1.D, we present estimates that have been "shrunk" with the Empirical Bayes procedure. This procedure adjusts for sampling error that may drive dispersion up; this appendix shows that our results are not sensitive to variable sampling error or the shrinkage procedure.

schools, in which dispersion is comparable across health groups, we find that the dispersion in school value-added is generally larger for unhealthy students. This pattern indicates that schools play a larger role in impacting academic outcomes for unhealthy students than for healthy students and that they may be more influential in generating test score gains among unhealthy students. In subsequent sections of this paper, we will explore how policies that aim to boost school effectiveness may be particularly salient for unhealthy students.

Table 1.6: Dispersion in Estimated School Effects

	Elementary school		Middle school		High school	
	Healthy students (1)	Unhealthy students (2)	Healthy students (3)	Unhealthy students (4)	Healthy students (5)	Unhealthy students (6)
	Outcome: ELA test scores					
Dispersion	0.115	0.151	0.148	0.157	0.136	0.163
	(0.003)	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)
Number of schools	764	764	496	496	525	525
	Outcome: Math test scores					
Dispersion	0.152	0.178	0.176	0.175	0.119	0.147
	(0.004)	(0.005)	(0.006)	(0.006)	(0.004)	(0.005)
Number of schools	763	763	497	497	525	525

Notes: This table reports dispersion of school effect estimates. Columns (1), (3), and (5) are for healthy students; Columns (2), (4), and (6) are for unhealthy students. We only include schools for which we estimate effects on *both* subgroups (healthy and unhealthy). Standard errors are calculated following Ahn and Fessler (2003), though one could also bootstrap the standard errors.

We find that when using ELA test scores as the dependent variable, dispersion is increasing from elementary to middle to high school. This indicates that the importance of schools to students' ELA test scores is increasing in student age, which could be related to the importance of writing and reading comprehension skills as students approach the college entry exam. On the other hand, with Math test scores as the dependent variable, dispersion is roughly decreasing from elementary to middle to high school (though monotonicity is less clear). This may be because Math is more closely related to cognitive or technical skills that students develop earlier in life (for example, people often say they "have a knack" for Math, but not for reading or writing).

With regard to subgroups, we find that dispersion in school effectiveness is larger among unhealthy students than among healthy students, almost across the board.[52] When using ELA test scores as the outcome, dispersion in school effectiveness is 31% larger among unhealthy students than among healthy students in elementary schools, 6% larger in middle schools, and 20% larger in high schools.[53] Larger dispersion indicates that there are more schools that are further above

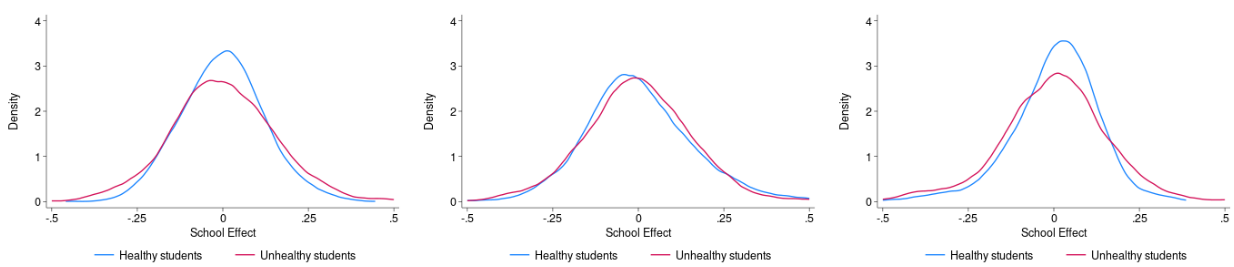
[52]The only exception is in middle schools, where dispersion is more comparable across the two subgroups.

[53]When using Math test scores as the outcome, dispersion in school effectiveness is 17% larger among unhealthy

(or below) the mean, which suggests that schools are more influential in determining students' test scores. The fact that dispersion in school effectiveness is consistently larger among unhealthy students than among healthy students indicates that schools have more power to boost test scores among unhealthy students than they do for healthy students, especially in elementary and high schools. We show in Appendix 1.D that these patterns of dispersion are robust to variable sampling error across schools and health groups of varying sizes.

Overall, our estimates of dispersion (0.12σ - 0.18σ) are in line with the past literature. Our estimates of 0.15σ - 0.18σ for middle schools are in line with what Angrist et al. (2017) find for middle schools, using a model that includes lagged test scores. Abdulkadiroğlu et al. (2020) find a standard deviation around 0.29σ for high schools (though they point out that their estimate is somewhat larger than other studies). The main caveat is that the estimates we provide in this section have not undergone the Empirical Bayes shrinkage procedure.[54] This means that the dispersion estimates that we present are smaller than in other studies if we were to compare to estimates that have not undergone the shrinkage procedure. There are a few of plausible explanations for this. The first is that we include a year fixed effect, which studies without panel data cannot include. We feel it is important to include the year fixed effect because factors that affect all schools but change over time should not be attributed to the effectiveness of individual schools. Second, we control more comprehensively for student heterogeneity with our health vector, which has removed additional variation that the school fixed effect otherwise picks up. In other words, our fixed effect estimates are less biased upward in magnitude (and our estimates of dispersion are also less biased upward). Lastly, related studies that use data from New York (e.g., Abdulkadiroğlu et al., 2020) or Boston (e.g., Angrist et al., 2017) may find more dispersion in value-added because there simply *is* more dispersion there than among Wisconsin's public schools.

Figure 1.8: Distributions of Estimated School Effects, by Health Group



(a) Elementary (b) Middle (c) High

Notes: Each graph shows distributions of de-meaned school effect estimates, separately by health group. For brevity we only include distributions that use ELA test scores as the outcome, though we find similar patterns in the regressions that use Math as the outcome.

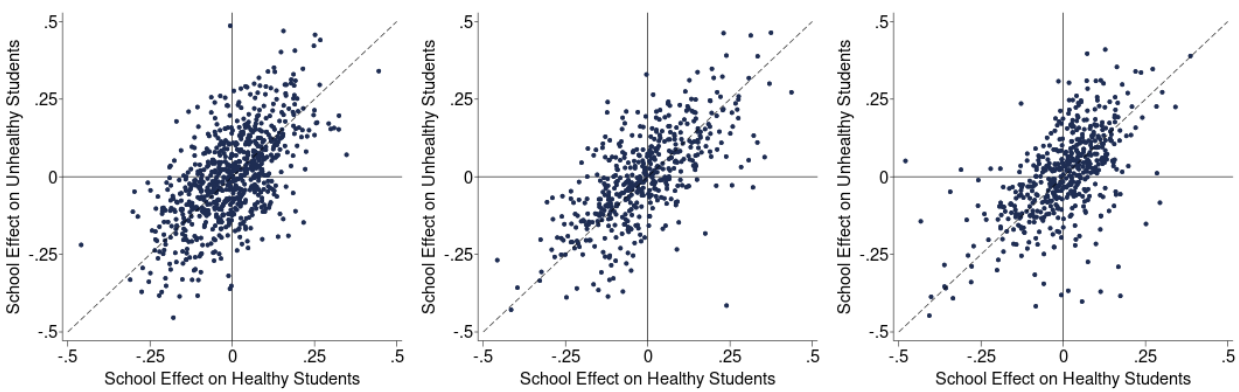
students than among healthy students in elementary schools and 24% larger in high schools; it is comparable across subgroups in middle schools.

[54]This is used to adjust for sampling error and is necessary when using value-added estimates on the right-hand side of subsequent regressions. The shrinkage procedure is not necessary in our context because our subsequent analysis places value-added on the left-hand side of a regression.

Figure 1.8 confirms the result from Table 1.6 that value-added dispersion (using ELA test scores as the outcome variable) is substantially larger among unhealthy students in elementary and high schools; even though dispersion is also larger among unhealthy students in middle schools, the difference is much smaller.[55] This is clear from the fact that the blue line (effect on healthy students) is more concentrated around the mean while the red line (effect on unhealthy students) has more density in the tails, particularly for elementary and high schools. There are several possible reasons for the findings being different in middle schools. It could be related to nonlinearity in human capital development over time, non-monotonicity of the importance of schools over time, or non-monotonicity in the interaction between health and school inputs to the development of human capital during childhood. We find that health is generally least severe among elementary school students while it is more severe among high school students; middle school students are often in the middle, indicating that there may be a monotonic production of health through time. With high schools containing the least healthy students, it may be the case that they have more policies in place specifically designed to mitigate the negative effects of poor health, which leads to more salient differences in school effectiveness across subgroups at the high school level. Middle schools have generally healthier students than high schools - albeit less healthy than elementary schools - so they are less pressed to give special attention to unhealthy students; treating healthy and unhealthy students the same leads to school effectiveness that varies much less across subgroups.

One immediate question arises from Figure 1.8: How does value-added compare across health groups within schools? Two possibilities come to mind. First, in order for a school to help one group of students, the school has less resources to be able to help the other group of students. For example, if a school expends resources toward helping its unhealthy students, maybe it has fewer

Figure 1.9: Correlation of the School Effects Across Health Groups



(a) Elementary (b) Middle (c) High

Notes: On the x-axis is the estimated school effect on healthy students and on the y-axis is the estimated school effect on unhealthy students, both using ELA test scores as the dependent variable in the value-added estimation. Each point represents a school. The 45-degree line is given as a reference to represent schools that have an equal effect on healthy and unhealthy students.

[55]Though we omit distributions that come from Math test scores for brevity, we find similar patterns.

remaining resources to help the healthier students. The second possibility is that schools that are better for one group are also better for the other group. In Figure 1.9, we plot school effect estimates across health groups, finding a strong positive correlation, which supports the second hypothesis. This is likely due to the fact that healthy and unhealthy students at the same school experience the same environment, same teachers, same peers, etc. Despite the strong positive correlation, we still find substantial variation in distance from the 45-degree line. This suggests that there could be differential match effects from school to school based on student health (i.e., some schools are more effective for healthy students while others are more effective for unhealthy students).

In general, schools above the 45-degree line generate larger gains to test scores (relative to the mean school) among unhealthy students than among healthy schools. Later in this section, we present results that show how school characteristics may drive some of the differential match effects; in the following section, we turn to a closer investigation of policies - specifically the hiring of homebound teachers and school nurses - designed for unhealthy students.

Table 1.7: Correlation in School Effect Estimates
Across Value-Added Outcomes and Health Groups

	Outcome: ELA Group: Healthy (1)	Math Healthy (2)	ELA Unhealthy (3)	Math Unhealthy (4)
	Elementary school			
Outcome: ELA, group: healthy	1.000			
Math; healthy	0.604	1.000		
ELA; unhealthy	0.592	0.462	1.000	
Math; unhealthy	0.508	0.706	0.565	1.000
	Middle school			
ELA; healthy	1.000			
Math; healthy	0.650	1.000		
ELA; unhealthy	0.686	0.509	1.000	
Math; unhealthy	0.521	0.763	0.543	1.000
	High school			
ELA; healthy	1.000			
Math; healthy	0.595	1.000		
ELA; unhealthy	0.578	0.456	1.000	
Math; unhealthy	0.397	0.586	0.609	1.000

Notes: This table reports correlations of estimates across value-added outcomes (ELA and Math) and across health groups (healthy and unhealthy). In each panel, the [2,1] and [4,3] entries represent correlations across outcomes, holding health group fixed; the [3,1] and [4,2] entries are correlations across health groups, holding outcome fixed; the [4,1] and [3,2] entries are correlations across estimates when both the outcome and health group switch. For the purpose of recovering these correlation coefficients, the value-added estimates have been normalized to be mean zero and standard deviation one by health group and school type.

We show in Table 1.7 that our school effect estimates are highly correlated not only across the two test score outcomes when holding health group fixed; they are also highly correlated also across health groups when holding outcome fixed. Interestingly, there is a moderate positive correlation across the school effect estimates even when allowing both test score outcome and health group to vary. The finding that our estimates are correlated across test score outcomes indicates robustness of the value-added model. The finding that our estimates are correlated across health groups indicates that more effective schools are generally more effective for each student subgroup. The correlation across subgroups is particularly strong when using Math test scores as the outcome. The finding that the estimates are correlated across both test score outcomes and health groups indicates that school resources that boost academic outcomes are effective for multiple facets of learning (different test subjects) and different types of students (healthy and unhealthy students).

Variance Decompositions

Decomposition 1: Value-Added Gap

In this section, we outline two simple variance decompositions to understand the contributions of subgroup-specific value-added to measures of effectiveness at the school level. In this subsection, we decompose variation in the school-level difference in value-added across health groups; in the next subsection, we decompose variation in overall school value-added. First, we define the school-level value-added "gap" to be the difference in school effects on healthy ($h = 1$) and unhealthy ($h = 0$) students:

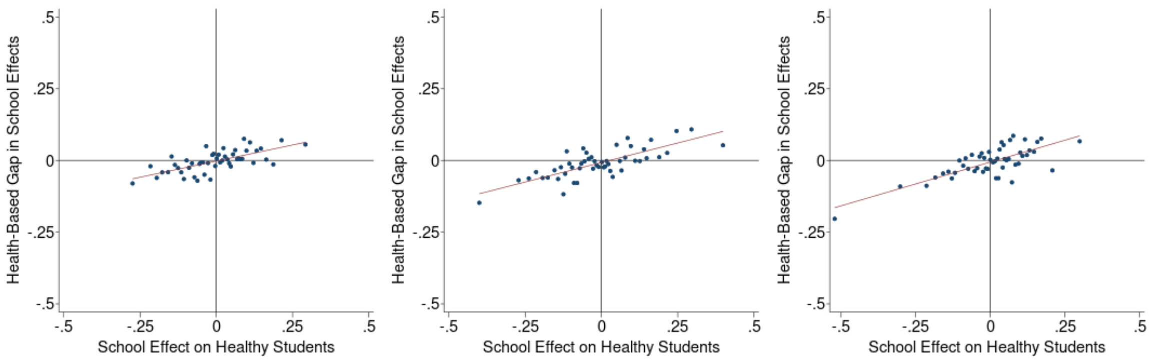
$$\text{Gap}_j = E_{j|h=1} - E_{j|h=0} \tag{1.8}$$

The value-added gap is simply a difference; it need not be positive in every school. A negative gap in school j would suggest that the effect of school j is actually more positive on unhealthy students than on healthy students (or less negative), relative to each respective mean school. We can decompose the variation in the school-level value-added gap to find the subgroup-specific value-added contributions, as follows:

$$Var(\text{Gap}_j) = Var(E_{j|h=1}) + Var(E_{j|h=0}) - 2Cov(E_{j|h=1}, E_{j|h=0}) \tag{1.9}$$

To motivate this decomposition of variance, we first uncover some patterns regarding the value-added gap. We plot the gap against the school effects on healthy students in Figure 1.10 and against the school effects on unhealthy students in Figure 1.11. An interesting pattern emerges: while there seems to be little relationship between the value-added gap and the school effect on the healthy group, there is a clear negative relationship between the gap and the school effect on the unhealthy group. In other words, we find that a marginal increase in how effective a school is for unhealthy students is associated with a clear decrease in the value-added gap between healthy and unhealthy students. However, a marginal increase in how effective a school is for healthy students is associated with a much smaller increase in the value-added gap. This might suggest that school policy that generates value to healthy students' test scores also boosts unhealthy students' test scores (learning-based policies like teacher training); school policy that generates value to unhealthy students' test scores might be designed specifically with those students in mind. In the next section, we explore this more, but first, we decompose variation in the value-added gap more explicitly; in addition, we decompose variation in overall school value-added.

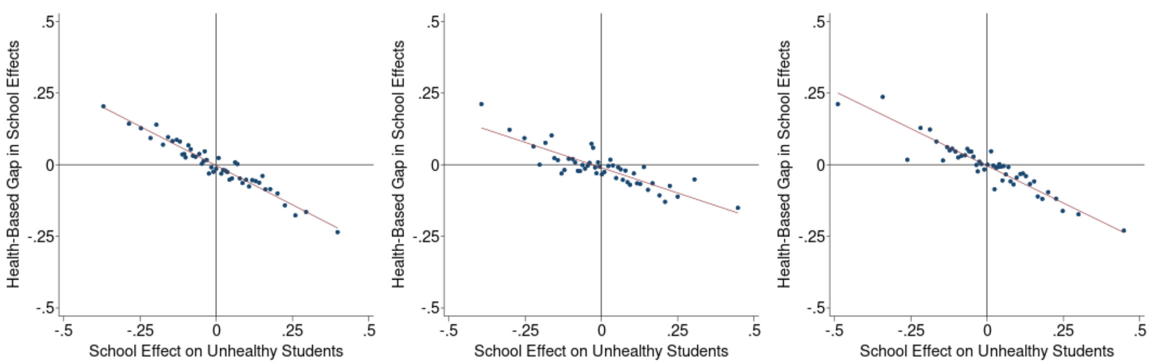
Figure 1.10: Explaining the Gap in School Effectiveness: Healthy Students



(a) Elementary (b) Middle (c) High

Notes: On the x-axis is the estimated school effect on healthy students using ELA test scores as the dependent variable, and on the y-axis is the value-added gap. We plot binscatters with 50 groups for the purpose of establishing a pattern on the distribution.

Figure 1.11: Explaining the Gap in School Effectiveness: Unhealthy Students



(a) Elementary (b) Middle (c) High

Notes: On the x-axis is the estimated school effect on unhealthy students using ELA test scores as the dependent variable, and on the y-axis is the value-added gap. We plot binscatters with 50 groups for the purpose of establishing a pattern on the distribution.

In Table 1.8 we decompose variation of the value-added gap across schools into components explained by the two subgroup-level school effects, as specified in Equation (1.9). With the exception of middle schools, we find that substantially more of the variation in the gap is explained by the school effect on unhealthy students than by the school effect on healthy students. In elementary and high schools, the school effect on unhealthy students explains as much as 50% more variation in the gap than the school effect on healthy students. Because the subgroup-specific school effects are (strongly) positively correlated with each other, as shown in Table 1.7, the covariance term explains a large portion of the gap. The findings from this decomposition help in explaining why there is such a weak relationship between the value-added gap and the school effect on healthy students, but a strong relationship between the gap and the school effect on unhealthy students.

Table 1.8: Subgroup-Level Contributions to Total Variance in the Value-Added Gap

	Elementary school	Middle school	High school
	(1)	(2)	(3)
	Outcome: ELA test scores		
Total variance of value-added gap	0.015	0.015	0.019
Variance of effect on healthy subgroup	0.013	0.022	0.018
Variance of effect on unhealthy subgroup	0.023	0.025	0.027
-2cov(effect on healthy, effect on unhealthy)	−0.021	−0.032	−0.026
	Outcome: Math test scores		
Total variance of value-added gap	0.017	0.015	0.015
Variance of effect on healthy subgroup	0.023	0.031	0.014
Variance of effect on unhealthy subgroup	0.032	0.031	0.022
-2cov(effect on healthy, effect on unhealthy)	−0.038	−0.047	−0.021

Notes: This table decomposes variation in the value-added gap across schools into components explained by subgroup-level school effects, as shown in Equation (1.9). The first row of each panel reports total variance in the gap (defined in Equation (1.8)); the second row reports the variance of school effects among healthy students; the third row reports the variance of school effects among unhealthy students; the last row shows the contribution of the covariance term.

Decomposition 2: Overall School Value-Added

Turning to the second variance decomposition, suppose that the overall effect of a school on its students' test scores is a weighted average of its effect on each health-based subgroup:

$$E_j = \underbrace{\left(\frac{N_{j|h=1}}{N_j}\right)}_{p_{j,1}} \underbrace{E_{j|h=1}}_{E_{j,1}} + \underbrace{\left(\frac{N_{j|h=0}}{N_j}\right)}_{p_{j,0}} \underbrace{E_{j|h=0}}_{E_{j,0}} \tag{1.10}$$

In Equation (1.10), the weights $p_{j,1} = \frac{N_{j|h=1}}{N_j}$ and $p_{j,0} = \frac{N_{j|h=0}}{N_j}$ are the proportions of school j's students that are healthy and unhealthy, respectively. In this regard, the health-based subgroup contributions to overall school value-added are weighted by the number of students in each health group. We can then decompose the variation in overall school value-added E_j as follows:

$$Var(E_j) = Var(p_{j,1}E_{j,1}) + Var(p_{j,0}E_{j,0}) + 2Cov(p_{j,1}E_{j,1}, p_{j,0}E_{j,0}) \tag{1.11}$$

The contribution of the school effect on health group $h = i$ to variation in overall school value-added can be calculated as $\frac{Var(E_{j|h=i})}{Var(E_j)}$, for $i = 0, 1$. However, we want to further disentangle the $E_{j,1}$ and $E_{j,0}$ from the health group proportions $p_{j,1}$ and $p_{j,0}$ to understand the unweighted contribution of each subgroup-level value-added and how the weights affect the contributions. To do this, we make use of the identity of the variance of a product of correlated objects, which (after some algebraic rearrangement) yields the following result that identifies disentangled weighted contributions of each subgroup effect:

$$Var(E_j) = w_1 Var(E_{j,1}) + w_0 Var(E_{j,0}) + r \tag{1.12}$$

In Equation (1.12), $w_1 = Var(p_{j,1}) + E\left[p_{j,1}\right]$ and $w_0 = Var(p_{j,0}) + E\left[p_{j,0}\right]$ serve as weights on the contributions of subgroup-specific value-added to overall school value-added. This allows us to understand how subgroup-level school effects might drive variation in overall school effects, and how forcefully these contributions are scaled down due to health group size. Lastly, the remainder r is defined in Equation (1.14) in the footnote below.[56]

In Table 1.9, we decompose variation of overall school effectiveness (which we define as a weighted average of the two subgroup-level value-added estimates) into components attributed to each subgroup; this variance decomposition is given in Equation (1.12). Similar to the previous result, we find that (with the exception of middle schools) a substantially larger contribution to overall school value-added comes from the variance of school effects on unhealthy students. However, because we construct the unhealthy group to contain about half as many students as the healthy group overall, the weight w_0 on the larger variance among unhealthy students is about half the magnitude of the weight w_1 on the smaller variance among healthy students. As a result, the weighted contribution of the school effect on health students on overall school value-added is larger than the weighted contribution of the school effect on unhealthy students.

Ultimately, the larger dispersion in school effects among unhealthy students is the primary driver of the results from this subsection: school effects among unhealthy students have larger contributions to the value-added gap and overall school value-added, at least prior to weighting by subgroup size.[57] These simple decomposition results in turn explain the un-intuitive result that the

[56]We can decompose the variance of a product as follows:

$$\begin{aligned} Var(p_{j,i}E_{j,i}) = {} & [Var(E_{j,i}) + E\,[E_{j,i}]]\,[Var(p_{j,i}) + E\,[p_{j,i}]] \\ & + Cov(p_{j,i}^2, E_{j,i}^2) - [Cov(p_{j,i}, E_{j,i}) + E\,[p_{j,i}]\,E\,[E_{j,i}]]^2 \end{aligned} \tag{1.13}$$

Through some algebra, we can separate out the $Var(E_{j,i})$ term in a way that it is attached only to a "weight" term of $Var(p_{j,i}) + E\,[p_{j,i}]$. The remainder r in Equation (1.12) is defined as follows:

$$\begin{aligned} r = {} & 2Cov(p_{j,1}E_{j,1}, p_{j,0}E_{j,0}) + \sum_{i=0}^{1} [E\,[E_{j,i}]\,[Var(p_{j,i}) + E\,[p_{j,i}]]] \\ & + \sum_{i=0}^{1} \left[Cov(p_{j,i}^2, E_{j,i}^2) - [Cov(p_{j,i}, E_{j,i}) + E\,[p_{j,i}]\,E\,[E_{j,i}]]^2\right] \end{aligned} \tag{1.14}$$

[57]It is also worth noting that these difference in dispersion across health groups is not due to differences in health

value-added gap is strongly related with the school effect on unhealthy students despite a weak relationship with the school effect on healthy students.

Table 1.9: Subgroup-Level Contributions to Variance in Overall School Value-Added

	Elementary school	Middle school	High school
	(1)	(2)	(3)
	Outcome: ELA test scores		
Variance of overall school value-added	0.013	0.019	0.016
Variance of effect on healthy	0.013	0.022	0.018
Weight on healthy variance (w_1)	0.700	0.684	0.674
Variance of effect on unhealthy	0.023	0.025	0.027
Weight on unhealthy variance (w_0)	0.309	0.331	0.380
Remaining terms (r)	−0.003	−0.004	−0.006
	Outcome: Math test scores		
Variance of overall school value-added	0.022	0.028	0.014
Variance of effect on healthy	0.023	0.031	0.014
Weight on healthy variance (w_1)	0.700	0.684	0.674
Variance of effect on unhealthy	0.032	0.031	0.022
Weight on unhealthy variance (w_0)	0.309	0.331	0.380
Remaining terms (r)	−0.004	−0.003	−0.004

Notes: This table decomposes variation in overall school value-added (as shown in Equation (1.12)), under the assumption that a weighted average of subgroup-specific school effects yields the overall school effect. The first row of each panel reports total variance in the overall school effect (defined in Equation (1.10)); the second row reports the variance of school effects among healthy students while the third row reports the weight on that contribution; the fourth row reports the variance of school effects among unhealthy students while the fifth row reports the weight on that contribution; the last row shows the contribution of second-order terms in the remainder r.

What Drives Effectiveness?

We now briefly turn to a descriptive exercise to understand the school characteristics that play a role in how effective a school is, and whether these characteristics vary by health subgroup. To this end, we project school effectiveness on several school characteristics (aggregated at the school level) separately by health group, as follows:

$$E_{j|h} = \beta_0^h + C_j'\beta_C^h + D_j'\beta_D^h + Q_j'\beta_Q^h + \beta_B^h B_j + \varepsilon_{jh} \tag{1.15}$$

C_j contains employee count variables: number of nurses at school j, homebound teachers, physical therapists, and psychologists. D_j contains demographic characteristics of teachers: percentage of teachers at school j that are White, Black, and female. Q contains what we label qualification

group size. We might be worried about larger sampling error among estimates for the unhealthy group; we show in Appendix 1.D that dispersion in school effectiveness is still larger among the unhealthy group even after shrinking.

variables; these are the percentage of teachers at school j with a graduate degree, and the average teacher experience in years. B_j contains demographic characteristics of students: percentage of students at school j that are White, Black, female, and enrolled in special education. In addition B_j contains the percentage of students at school j that are unhealthy. Since our measures of school effectiveness do not vary over time, we average each variable on the right-hand side over all observed years 2009-2018. In this regard, the only variation we are working with is at the school level (and the health group level because we run separate regressions for each health group). We also scale each right-hand side variable such that coefficients can be interpreted as the association between a one standard deviation increase in that variable and school effectiveness. In addition, we run regressions separately for each school type, because we expect school characteristics and their influence on effectiveness to vary across school types.

Rather than present all coefficient estimates from Equation (1.15) here, we leave them in Appendix 1.C (Tables 1C.5, 1C.6, and 1C.7). We instead show in Table 1.10 the contributions of school characteristics to total variation in school effectiveness. We keeps things aggregated at the vector level (rather than breaking out individual characteristics) for simplicity. In elementary schools, no vector of characteristics stands out as being particularly crucial in explaining variation in school effectiveness. The same can be said when breaking the vectors into individual characteristics (shown in Appendix 1.C): coefficients on nearly all school characteristics are both economically insignificant and statistically insignificant. The only exception is that elementary schools with more female teachers tend to have very slightly higher value-added on both subgroups. In middle schools we find higher contributions of teacher demographics and qualifications than in elementary schools. In addition, we find that a one standard deviation increase in average teacher experience in middle schools is associated with an increase in effectiveness of around 0.03 test score standard deviations; this increase is stable across healthy and unhealthy student groups.

In high schools, a few interesting patterns emerge. First, student demographics explain substantially more variation in school effectiveness than any other vector of school characteristics; this finding is robust to test score outcome and health group. This would suggest that the sorting of students into schools is a notable factor in school effectiveness, even after comprehensively controlling for sorting in the value-added model. We find that much of explanatory power of student demographics is due to student health, further indicating the importance of controlling for health in models of value-added. Second, increases in school nurses are associated with decreases in school effectiveness on healthy students, but there is little association between school nurses and school effectiveness on unhealthy students. This suggests that schools may be resource constrained and that when they hire nurses (who are intended to help unhealthy students) they are forced to divert more resources away from healthy students, which results in decreased value-added among those students. Third, a one standard deviation increase in the percentage of the student body that is unhealthy is associated with substantial drops in value-added. Furthermore, the drop among healthy students (0.07σ) is notably larger than the drop among unhealthy students (0.05σ). While this supports the hypothesis that schools with more unhealthy students may divert more resources away from healthy students, it also suggests that policies that improve student health could have positive spillover effects on the productivity of high school inputs. In other words, if policymakers could reduce the number of unhealthy students at a school, the school may become more effective for both the healthy and unhealthy students.

Table 1.10: Variation in School Effectiveness Explained by Observables

	Elementary school		Middle school		High school	
	Healthy students (1)	Unhealthy students (2)	Healthy students (3)	Unhealthy students (4)	Healthy students (5)	Unhealthy students (6)
	Outcome: ELA test scores					
Total variation explained by:						
Employee counts (C_j)	0.041	0.013	0.039	0.027	0.037	0.027
	(0.016)	(0.006)	(0.012)	(0.009)	(0.015)	(0.010)
Teacher demographics (D_j)	0.031	0.027	0.059	0.045	0.031	0.077
	(0.009)	(0.008)	(0.018)	(0.016)	(0.011)	(0.030)
Teacher qualifications (Q_j)	0.012	0.015	0.038	0.046	0.019	0.021
	(0.005)	(0.006)	(0.016)	(0.018)	(0.008)	(0.008)
Student demographics (B_j)	0.050	0.037	0.049	0.050	0.178	0.149
	(0.012)	(0.012)	(0.014)	(0.016)	(0.039)	(0.029)
Overall R-squared	0.134	0.092	0.185	0.168	0.265	0.274
Number of schools	762	762	477	477	517	517
	Outcome: Math test scores					
Total variation explained by:						
Employee counts (C_j)	0.025	0.021	0.030	0.034	0.030	0.015
	(0.009)	(0.008)	(0.009)	(0.009)	(0.013)	(0.006)
Teacher demographics (D_j)	0.036	0.035	0.041	0.047	0.045	0.082
	(0.011)	(0.009)	(0.015)	(0.014)	(0.013)	(0.030)
Teacher qualifications (Q_j)	0.016	0.014	0.040	0.035	0.030	0.024
	(0.007)	(0.006)	(0.017)	(0.013)	(0.012)	(0.010)
Student demographics (B_j)	0.040	0.045	0.070	0.086	0.200	0.146
	(0.012)	(0.012)	(0.021)	(0.022)	(0.030)	(0.025)
Overall R-squared	0.117	0.115	0.181	0.202	0.305	0.267
Number of schools	762	762	477	477	517	517

Notes: This table reports the variation in school effectiveness that is explained by factors relating to employee counts (C_j), teacher demographics (D_j), teacher qualifications (Q_j), and student demographics (B_j). These are the vectors used in Equation (1.15); for simplicity, we only present aggregate variation accounted for by each vector (rather than breakdowns by each individual characteristic). Each element of the table represents the portion of the overall R-squared that can be attributed to a particular category of school characteristics. Bootstrapped standard errors (using 500 repetitions) are in parenthesis.

1.6 Evaluating School Policy

Two-Way Fixed Effects Model

Now that we have established clear heterogeneity in school effectiveness based on student health, we investigate how a few school-level policies directly impact students of varying health. Specifically, we look at how the hiring of homebound teachers[58] and nurses[59] affect students' cognitive outcomes (ELA and Math test scores) and non-cognitive outcomes (absence rate, special education enrollment, and time taken to graduate from high school). Since we know which schools hire nurses and homebound teachers but not how students utilize these resources, our analysis can be thought of as an intent-to-treat; we estimate the effects of nurse and homebound teacher hiring, but not the true effect of nurses and homebound teachers on the students they serve.

Intuitively, the effects of school policy on students' cognitive and non-cognitive outcomes pass through the channel of school effectiveness (schools choose policy and therefore the impacts of these policies are part of what we think of as school quality). To match our econometric model to this economic intuition, we now allow school effectiveness to vary from year to year (i.e., we estimate θ_{jht} in the first-stage school value-added model instead of θ_{jh}),[60] estimating the effects of school policy in T_{jt} on student outcomes in Y_{it} with the following two-stage model:

$$Y_{it} = \theta_{jht} + X'_{it}\beta_X + H'_{it}\beta_H + \omega_t + \varepsilon_{it} \tag{1.16}$$

$$\theta_{jt|h} = \beta_0^h + \beta_T^h T_{jt} + \beta_S^h S_{d(j),t} + N'_{jt}\beta_N^h + \delta_j^h + \omega_t^h + \varepsilon_{iht} \tag{1.17}$$

Before discussing each variable contained in the model, we first describe the model at a high level to solidify the economic intuition. Equation (1.16) represents the first stage: here, we re-estimate the school fixed effects in a way that they not only vary across health-based subgroups, but also across time. Equation (1.17) represents the second stage of the model: here, we estimate the impacts of school policy on school effectiveness with a two-way fixed effects design (which includes a school fixed effect and academic year fixed effect). Specifically, we aim to understand how the hiring of homebound teachers and school nurses impact school effectiveness, paying specific attention to health-based student subgroups.

Econometrically, it is simpler for us to estimate the two-stage model described above in one step (i.e., plug Equation (1.17) into Equation (1.16)), as written below. In this regard, we estimate the effects of homebound teachers and school nurses (school policy) on students' cognitive and non-cognitive outcomes, under the intuition that these effects are encompassed by what we think of as school effectiveness. For homebound teachers, we run separate regressions for healthy and

[58] Recall that when students with disabilities or severe illnesses are unable to attend school in person, homebound teachers - employed by schools or districts - travel to students' homes to provide instruction.

[59] During our analysis window of 2009-2018, Wisconsin has no mandate on school nurses.

[60] Our baseline specification does not allow time-variation in school effectiveness to increase cell size and decrease sampling error in the fixed effect estimates.

unhealthy students to estimate their effects on these groups individually; we portray this by denoting outcomes as being conditional on health group h. Nurses are endogenous to student health, which complicates things if we were to try to estimate separate effects on healthy and unhealthy students; for this reason, we estimate nurses' effects on student outcomes unconditional on health.

The econometric model for the effect of homebound teacher hiring is written as follows:

$$Y_{it|h} = \beta_0^h + \beta_T^{h,HB} T_{jt}^{HB} + \beta_S^h S_{d(j),t} + N_{jt}'\beta_N^h + X_{it}'\beta_X^h + H_{it}'\beta_H^h + \delta_j^h + \omega_t^h + \varepsilon_{iht} \tag{1.18}$$

The econometric model for the effect of nurse hiring can be written as follows:

$$Y_{it} = \beta_0 + \beta_T^{RN} T_{jt}^{RN} + \beta_S S_{d(j),t} + N_{jt}'\beta_N + X_{it}'\beta_X + H_{it}'\beta_H + \delta_j + \omega_t + \varepsilon_{it} \tag{1.19}$$

As mentioned, we include in Y_{it} ELA test scores, Math test scores, absence rate, special education enrollment, and time taken to graduate as outcome variables. In addition, because school nurses are hired with health in mind, we also investigate how they impact student health. T_{jt}^{HB} in Equation (1.18) is a variable for the number of homebound teachers that students at school j have access to in year t (HB for homebound teacher); similarly, T_{jt}^{RN} in Equation (1.19) is a variable for the number of nurses that students at school j have access to in year t (RN for registered nurse). We refer to these two variables interchangeably as school policies and treatments. The β_T coefficients on the treatment variables in each equation are the parameters that we are interested in estimating; they represent the causal intent-to-treat effects of homebound teacher hiring and school nurse hiring on a range of student outcomes.

Many homebound teachers and nurses are employed by school districts (rather than schools), and they serve multiple schools in the district. In calculating the number of nurses that work at a particular school, we account for both nurses employed by the school and the district with the formula outlined in Equation (1.20). The homebound teacher formula is analogously defined. In words, the total nurse count at school j is the sum of two things: (1) the average number of nurses assigned to district $d(j)$ per school j in the district, and (2) the number of nurses assigned to school j. This is done separately by year.[61]

$$[\text{Nurse count}]_{jt} = \frac{[\text{\# district nurses}]_{d(j)t}}{\left[\sum_j \mathbb{1}\left[j \in d(j)\right]\right]_t} + [\text{\# school nurses}]_{jt} \tag{1.20}$$

Though the employment of nurses and homebound teachers are endogenous choices made by schools, these are not grounds for throwing out a causal interpretation. To deal with this

[61]This calculation relies on the assumption that district nurses spend equal time at each of the schools in the district, at least in expectation. Realistically, there is a large variation in school size within districts, and district nurses likely end up spending more time at the larger schools; however, these nurses are responsible for students at all of the schools within the district, so it is fair to assume that they expect to spend a non-zero amount of time at each of the schools, though it may not be exactly equal.

endogeneity, we control for the school-side factors that are most likely to determine policy. District spending $S_{d(j),t}$ proxies for resources available at the district level and accounts for any differences in resources that are available to different schools. Total school enrollment and percentage of students in a school that are unhealthy are included in N_{jt}, and are both important in accounting for other factors that may determine the hiring of nurses and homebound teachers. Total school enrollment is necessary because larger schools (especially those in larger districts) have more resources available and are able to hire more staff. The percentage of students in a school that are unhealthy is important because it indicates a school's need to pour more resources into aiding unhealthy students. Other factors influencing school's hiring decisions - such as an active community that advocates for its unhealthy students - will be picked up by the school fixed effect. Controlling for these factors that are endogenous to nurse and homebound teacher employment is important in removing any omitted variable bias that would otherwise be picked up by β_T^{HB} and β_T^{RN}.

X_{it} and H_{it} are similar to the vectors from Equation (1.4) described in Section 1.5, though some covariates have been removed either because they are now outcomes of interest or to prevent over-controlling in the model. X_{it} includes past test scores (in a fully interacted cubic polynomial in past ELA and Math test scores), subsidized lunch status, cumulative years eligible for subsidized lunch, race/ethnicity, gender, and an indicator for limited English proficiency; H_{it} includes the predicted health-related absence rate, lagged and twice-lagged predictions, number of ER visits, and number of Medicaid claims.[62] We include a school fixed effect δ_j to soak up any school-level factors that are constant across time and an academic year fixed effect ω_t to soak up any statewide policies during the analysis window of 2009-2018. Consistent with the value-added estimation, the model is estimated separately by school type (elementary, middle, and high school).

In order to interpret the β_T^{HB} and β_T^{RN} coefficients in a causal way, we make a few crucial identifying assumptions, following from de Chaisemartin and D'Haultfoeuille (2020). First, we assume that the employment of nurses and homebound teachers is not based on outcomes at baseline. Nurses are likely to be hired more often when students are unhealthy but this is not a concern because we estimate effects of nurses on students unconditional on health. Furthermore, while we do look at the effects of nurses on health, it is not a primary outcome and serves more to understand the mechanisms through which nurses impact absence rates and test scores; the hiring of nurses is not dependent on our other outcomes. Existing teachers gaining homebound certification as well as the hiring of new homebound teachers is due primarily to special education enrollment.[63] In this regard, the specification looking at the effect of homebound teachers on special education enrollment violates the treatment exogeneity assumption, and so we interpret this small subset of the results as correlational rather than causal.[64] We find no obvious reasons for concern

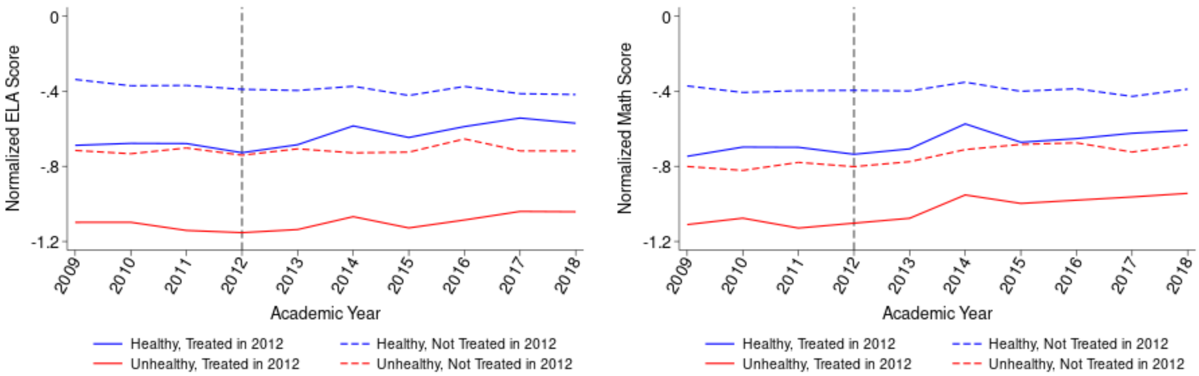
[62]Diagnosis codes have been omitted from this econometric model because many are highly correlated with each other and make coefficient interpretations more difficult.

[63]Recall that qualification for providing at-home instruction requires teachers to have a bachelor's degree and special education certification.

[64]Teachers may also be incentivized to gain homebound qualification if many students are absent from class, though we are less concerned about this. Given teachers' time constraints, it is unlikely for a teacher of a non-special education class to be able to separately devote time to teaching students in school as well as students that are absent. The more

using test scores as the outcome. Homebound teachers are not hired directly because students have low test scores; they are hired based on non-cognitive factors and district budget.

Figure 1.12: Parallel Trends in Test Scores (Homebound Teacher Policy)



(a) ELA Test Scores

(b) Math Test Scores

Notes: Each graph shows trends in test scores among high school students; schools that hired homebound teachers prior to 2012 are already treated and not included. The solid blue line is for healthy students in schools with a new homebound teacher in 2012; the dashed blue line is for healthy students in schools without a new homebound teacher in 2012; the solid red line is for unhealthy students in schools with a new homebound teacher in 2012; the dashed red line is for unhealthy students in schools without a new homebound teacher in 2012. For parallel trends, we compare the solid and dashed lines of a single color.

Second, we assume that expected outcomes in the absence of treatment would follow the same path for students that are and are not actually treated (this is commonly referred to as the parallel trends assumption). In the context of the unhealthy student group and the homebound teacher policy, since we estimate Equation (1.18) separately by health group, we also need parallel trends to hold separately for healthy and unhealthy students. We assume outcome trends among unhealthy students at schools with at least one homebound teacher to follow the same trajectory - in the absence of treatment - as outcome trends among unhealthy students at schools with no homebound teachers.[65] Since we have several years (as opposed to the standard two-period model) with tremendous variation in treatment timing, there are two caveats to the parallel trends assumption. First, the assumption must hold in each year for the students that are newly treated. And second, the assumption must hold when comparing newly treated students to never-treated students and to not-yet-treated students. In practice, the satisfaction of this assumption is illustrated with parallel trends prior to treatment. In Figure 1.12, we compare trends in ELA and Math test scores - separately by health group - for high school students with access to a homebound teacher in 2012 against high school students without access in 2012 (the untreated group includes both the never-treated students and the not-yet treated students); this is the year in which we see the largest increase in homebound

likely situation is that the school's teachers that are already homebound certified provide the homebound instruction. Therefore, we are not worried about absence as an outcome.

[65]We make analogous assumptions regarding the healthy student group. With nurses as the school policy, we assume parallel trends in outcomes, unconditional on health group.

teacher employment (variation in homebound teacher employment and nurse employment can be found in Figure 1C.5 in Appendix 1.C). We present evidence of parallel trends for the other outcomes as well as for school nurse policies in Appendix 1.C, but for brevity, we omit graphs for the other school types and years.[66] Reassuringly, we find pre-trends to be parallel for ELA and Math test scores as well as absence rates and graduation timing. Trends in special education enrollment are less clearly parallel, but we have already cautioned against interpretation of effects on special education enrollment as causal based on a breakdown of the treatment exogeneity assumption.

To allow for the ending of treatment (in our context, schools decreasing their nurse or homebound teacher employment), de Chaisemartin and D'Haultfoeuille (2020) find that two additional assumptions are needed. First, we assume that the termination of employment of nurses and homebound teachers is not based on students' outcomes. We argue that decreases in nurse and homebound teacher employment are related to the district budget above all else, which we control for. Second, we assume that expected treated outcomes would follow the same path for students that do and do not actually end treatment.[67] Keeping the context of the unhealthy student group and the homebound teacher policy, we assume outcome trends among unhealthy students at schools that cut ties with homebound teachers to follow the same trajectory - if treatment did not end - as outcome trends among unhealthy students at schools that keep their homebound teachers. This is not very common in the data, so we omit a presentation of the relevant parallel trend graphs for brevity. Under the four assumptions outlined in this section, we interpret β_T as a causal average treatment effect of employee hiring on the treated.

The Effects of Homebound Teachers and Nurses

In Table 1.11 we present results from the two-way fixed effects design specified in Equation (1.18) to estimate the impact of homebound teachers on students' test scores. In elementary schools, we find null effects of homebound teacher presence. We suspect this to be a consequence of lower utilization of homebound teachers among elementary school students; this could be related to the fact that course subjects in elementary schools are introductory and involve little specialization, making it easier for parents to help their kids in the event of absence. In middle schools, we find imprecisely estimated effects of homebound teacher hires on Math scores, but statistically significant positive effects on ELA scores. These findings hold both for healthy and unhealthy students. Towards the end of middle school, math courses become more tracked while there is little to no tracking in English courses. It may be the case that during this transition into a fully tracked system, in-school learning in Math does not translate to at-home instruction.

[66] With three types of schools (elementary, middle, and high school), 8 years of feasible treatment years where both a pre-treatment and post-treatment period is observed (2010-2017), two policies (homebound teachers and nurses), and five primary outcomes of interest (Math scores, ELA scores, absence rate, special education enrollment, and time taken to graduate), there are 360 unique trend comparisons we could present (since comparisons for nurse policies are not at the health-based subgroup-level).

[67] This is analogous to the aforementioned parallel trends assumption, but for treated students leading up to the end of treatment rather than for treated students leading up to the start of treatment.

Table 1.11: Effect of Homebound Teachers on Test Scores (Equation (1.18))

	Effect on Math test scores		Effect on ELA test scores	
	Healthy students (1)	Unhealthy students (2)	Healthy students (3)	Unhealthy students (4)
	Elementary school students			
Homebound teacher effect	−0.016	−0.003	0.008	−0.002
	(0.021)	(0.031)	(0.020)	(0.029)
Mean test score	−0.352	−0.571	−0.361	−0.562
R-squared	0.553	0.532	0.596	0.574
Student-year observations	95,494	43,780	95,537	43,816
	Middle school students			
Homebound teacher effect	0.111	−0.112	0.246	0.173
	(0.059)	(0.072)	(0.055)	(0.067)
Mean test score	−0.395	−0.728	−0.376	−0.670
R-squared	0.535	0.510	0.586	0.562
Student-year observations	96,839	44,959	96,908	45,047
	High school students			
Homebound teacher effect	0.012	0.065	0.027	0.054
	(0.014)	(0.019)	(0.015)	(0.020)
Mean test score	−0.454	−0.815	−0.433	−0.802
R-squared	0.638	0.557	0.627	0.578
Student-year observations	103,155	43,125	102,938	42,903
School fixed effects	Y	Y	Y	Y
Academic year fixed effects	Y	Y	Y	Y

Notes: This table presents results from the two-way fixed effect regression specified in Equation (1.18) with test scores on the left-hand side. All regressions include school and academic year fixed effects, along with controls for student heterogeneity in demographics, health, and ability (past test scores), total school enrollment, percent unhealthy, and district-level spending per pupil. Mean test scores by school type and health group are provided to understand the effect of the homebound teachers in the context of pre-existing differences in test scores.

In high schools, we find null effects of homebound teacher hires on healthy students but statistically significant positive effects on unhealthy students, regardless of test subject (Math or ELA). In particular, the marginal effect of hiring a homebound teacher on the average unhealthy high school student is a 0.065 standard deviation (SD) boost to Math test scores and a 0.054 SD boost to ELA test scores. The null effects on healthy students are expected, since homebound teachers are less likely to come into contact with healthy students (i.e., healthy students miss less school and therefore have less of an opportunity to utilize homebound teachers). The positive effects on unhealthy students are also unsurprising: after all, students that miss time due to illness are the primary target group of homebound teachers. In addition, homebound teachers are most often aimed at helping high school students, since this is where poor health and absence have the largest impacts on more specialized cognitive skills.

Table 1.12: Effect of Homebound Teachers on Non-Cognitive Outcomes (Equation (1.18))

	Effect on absence rate		Effect on special education enrollment		Effect on time taken to graduate	
	Healthy students (1)	Unhealthy students (2)	Healthy students (3)	Unhealthy students (4)	Healthy students (5)	Unhealthy students (6)
			Elementary school students			
Homebound teacher effect	−0.038	−0.277	0.001	−0.016	-	-
	(0.155)	(0.254)	(0.011)	(0.017)		
R-squared	0.106	0.232	0.252	0.299	-	-
Student-year observations	95,776	44,102	95,875	44,175	-	-
			Middle school students			
Homebound teacher effect	−1.242	0.371	−0.040	−0.032	-	-
	(0.554)	(0.897)	(0.040)	(0.038)		
R-squared	0.133	0.370	0.171	0.325	-	-
Student-year observations	97,755	46,122	98,801	46,183	-	-
			High school students			
Homebound teacher effect	−0.438	−0.901	−0.007	−0.022	−0.009	−0.025
	(0.259)	(0.500)	(0.009)	(0.012)	(0.008)	(0.015)
R-squared	0.193	0.346	0.309	0.306	0.063	0.160
Student-year observations	105,601	49,233	105,743	49,391	92,307	36,063
School fixed effects	Y	Y	Y	Y	Y	Y
Academic year fixed effects	Y	Y	Y	Y	Y	Y

Notes: This table presents results from the two-way fixed effect regression specified in Equation (1.18) with non-cognitive outcomes on the left-hand side. All regressions include school and academic year fixed effects, along with controls for student heterogeneity in demographics, health, and ability (past test scores), total school enrollment, percent unhealthy, and district-level spending per pupil.

Table 1.12 presents results from the two-way fixed effects design with homebound teachers as the treatment, using non-cognitive outcomes as the dependent variable. Our null findings among elementary school students extends to absence rates and special education enrollment. Homebound teachers appear to decrease the absence rate among middle school students, though not among unhealthy students. Among high school students, we find that homebound teachers decrease absences among both healthy and unhealthy students; perhaps absence is counted differently when homebound intervention occurs. Among healthy high school students, we find null effects of homebound teachers on special education enrollment and time taken to graduate (confirming the hypothesis that these students have little contact with homebound teachers). Among unhealthy high school students, we estimate small decreases in special education enrollment as well as time taken to graduate, which indicates success of the homebound education system at improving outcomes of unhealthy students beyond standardized exam performance. The results here should be taken as a positive signal by administrators implementing policies for unhealthy students.

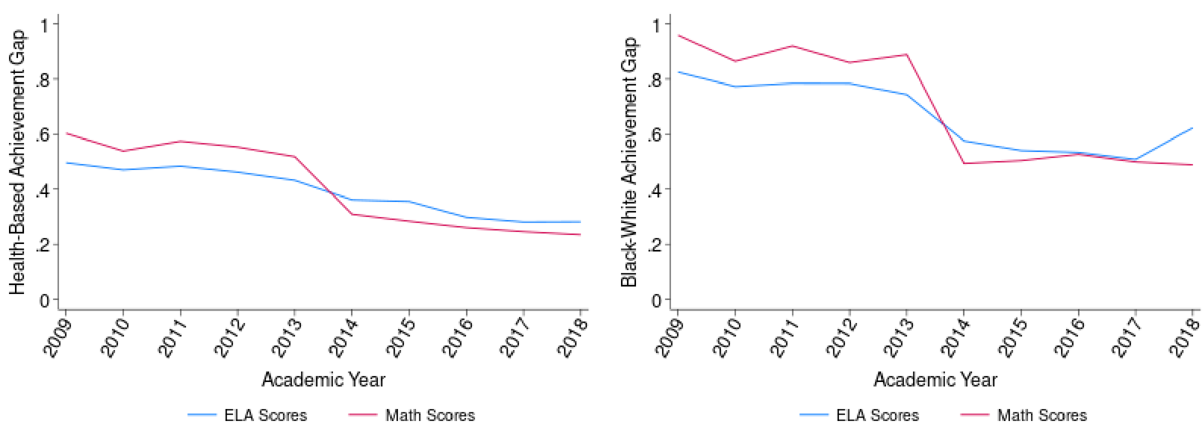
Table 1.13: Effect of Nurses on Cognitive and Non-Cognitive Outcomes (Equation (1.19))

	Math score (1)	ELA score (2)	Absence rate (3)	Special ed. enrollment (4)	Time taken to graduate (5)	Student health (6)
			Elementary school students			
Nurse effect	−0.007	−0.0004	−0.090	−0.017	-	0.019
	(0.012)	(0.011)	(0.092)	(0.006)		(0.021)
R-squared	0.565	0.604	0.120	0.271	-	0.216
Student-year observations	184,778	184,883	185,520	185,778	-	185,778
			Middle school students			
Nurse effect	0.002	−0.004	0.076	−0.001	-	−0.001
	(0.003)	(0.003)	(0.032)	(0.002)		(0.011)
R-squared	0.552	0.594	0.154	0.231	-	0.217
Student-year observations	191,539	191,722	194,170	194,404	-	194,404
			High school students			
Nurse effect	0.0001	0.001	0.623	−0.008	0.001	−0.003
	(0.004)	(0.005)	(0.093)	(0.003)	(0.003)	(0.027)
R-squared	0.639	0.628	0.264	0.305	0.095	0.355
Student-year observations	196,853	196,315	206,705	207,098	175,484	207,098
School fixed effects	Y	Y	Y	Y	Y	Y
Academic year fixed effects	Y	Y	Y	Y	Y	Y

Notes: This table presents results from the two-way fixed effect regression specified in Equation (1.19) to understand the effect of nurses on various outcomes. All regressions include school and academic year fixed effects, along with controls for student heterogeneity in demographics, health, and ability (past test scores), total school enrollment, percent unhealthy, and district-level spending per pupil.

Turning to school nurses, we find null effects for the most part, as shown in Table 1.13. The exceptions are statistically significant increases in the absence rate among middle and high school students and a decrease in special education enrollment among elementary school students. Nurses likely increase absences because students that feel unwell (whether or not they are actually sick or unhealthy) visit the nurse during school and the nurse sends them home. Nurses may decrease special education enrollment in elementary schools if they help in identifying disabilities for which the negative effects can be mitigated if properly addressed (ADHD, for example); furthermore, it may be early enough for health intervention to be key in keeping students that do not need special education out of special education. We find an imprecise positive impact of nurses on elementary school students' health, so more work is needed to fully understand these patterns. Part of the explanation for overwhelmingly null findings among nurses could be a lack of variation in nurse employment over time. Figure 1C.5 shows that while most schools employ at least one part-time nurse (roughly 75-90% in any given year and school type), there is little variation over time.

Figure 1.13: Achievement Gaps Across Time



(a) Health-based achievement gaps (b) Black-White achievement gaps

Notes: Panel (a) shows achievement gaps between unhealthy and healthy high school students from 2009-2018 (a positive number indicates higher scores among healthy students). Panel (b) shows achievement gaps between Black and White high school students. We focus on high schools due to the homebound teacher distribution across school types. In 2014, Wisconsin replaced the WKCE exam for high school students with the ACT, which removed some of the inherent bias toward students from higher-resource households. The two-way fixed effect models are robust to removal of the pre-ACT academic years.

We find the average health-based achievement gap across years among lower-income Medicaid-enrolled high school students to be around 0.4 standard deviations (both for ELA and for Math), as shown in Table 1.11.[68] We present trends in the high school health-based achievement gap in Figure 1.13 along with trends in the Black-White achievement gap, a longstanding phenomenon that has been studied extensively but is still not fully understood. Graphically, we find that the statewide health-based achievement gap is roughly half of the race-based achievement gap; regression analysis suggests that health directly explains 5-6% of race-based achievement gaps within high schools. With an estimated effect of 0.065 Math test score standard deviations on unhealthy high school students and a null effect on healthy high school students, hiring homebound teachers closes health-

[68]On top of the pre-existing achievement gaps around 1 standard deviation between students in our sample and non-Medicaid students that are not in our sample, as shown in Table 1.1.

based achievement gaps by around 16%.[69] Similarly, homebound teachers directly close ELA gaps by 7-14%. These results imply a reduction in longstanding race-based achievement gaps as well. Importantly, these gaps are not shrinking by way of harm to the healthy students (i.e., unhealthy students benefit but at the expense of healthy students). Instead, the primary driver is a boost to unhealthy students' test scores while healthy students' scores remain stable, which indicates that homebound teacher policies result in an increase in student welfare overall.

1.7 Discussion

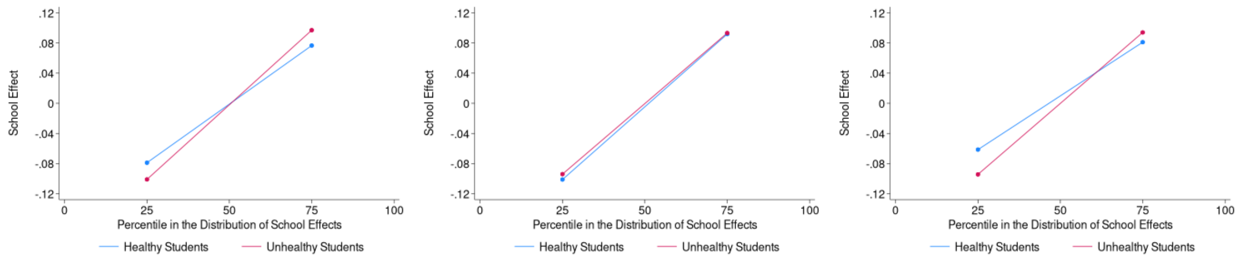
There are several interesting findings of this paper that need to be unpacked. First, we find that in general, the dispersion in school effects on unhealthy students is substantially larger than the dispersion among healthy students. This would suggest that match effects may be more salient among unhealthy students and that schools have more potential to influence outcomes among these students. We also find that the schools that are better for unhealthy students (the schools that are the highest value-added) are also typically better for healthy students, and that the correlation is quite strong. This would suggest that students of varying underlying health that attend the same school - and are exposed to the same teachers, the same peers, the same curriculum, etc. - are impacted by the school in similar ways.

In general, the value-added gap - which we define as the school effect on healthy students minus the school effect on unhealthy students - is small, which is driven by the high positive correlation between the two effects. Interestingly, much of the variation in the gap is driven by the school effect on unhealthy students. This means that a marginal increase in the effect of a school on unhealthy students can go a long way in boosting outcomes among unhealthy students without negatively impacting healthy students. Consider the following thought experiment: denoting an ineffective school by the 25th percentile and an effective school by the 75th percentile, what would happen if we reallocate students from an ineffective school to an effective school? Figure 1.14 shows that the boost in school effectiveness would be substantially larger (over 30% larger) for an unhealthy student than for a healthy student in elementary and high school (while the gain is similar across health groups for students in middle school). In fact, the increase in school value-added nearly wipes away enormous pre-existing achievement gaps between healthy and unhealthy students.

Though this provides suggestive evidence of potential gains from the reallocation of unhealthy students, these issues are more complex in practice and need to be studied more carefully in future work. With schools and districts often facing binding resource constraints, how would a policymaker reallocate students from schools that are not effective to schools that are, without hurting the students that already attend the more effective schools? This opens the door for several interesting questions regarding school effectiveness, reallocation of students, and dynamics in health (i.e., students that become healthy or unhealthy).

[69] Even if we remove the imprecisely estimated effect on healthy students, we estimate that homebound teachers close health-based achievement gaps in high schools by 13%.

Figure 1.14: Reallocating Students from an Ineffective School to an Effective School



(a) Elementary (b) Middle (c) High

Notes: These graphs depict the change in school value-added on healthy and unhealthy students if we reallocate them from a school at the 25^{th} percentile (ineffective) to a school at the 75^{th} percentile (effective). Value-added is estimated from ELA test scores, though patterns are similar for Math test scores.

An alternative to student reallocation is policy intended to boost value-added among schools that are less effective. In this paper we begin to study the role of nurses and homebound teachers, both of which have the primary purpose of aiding unhealthy students (students that feel unwell in the case of nurses, and students that miss school time due to poor health in the case of homebound teachers). We find that nurses have little effect on school value-added or students' academic outcomes, but that they may actually improve health; because there is no mandate on school nurses in Wisconsin, this may be an extremely important area for future research, as it may have several policy implications. We uncover suggestive evidence that homebound teachers are important in improving school effectiveness - particularly among unhealthy students - and that they also boost test scores among unhealthy middle and high school students. More research is needed to fully understand the importance of a homebound education system. Additional work on school policy in the realm of student health has the potential to help administrators understand how to boost school effectiveness for unhealthy students and improve academic outcomes among these students.

1.8 Conclusion

Through a combination of machine learning and value-added techniques, this paper examines complementarities between school and health inputs to the production of human capital. We use a random forest to measure student health, mapping diagnosis codes to student's absence rates in school to estimate students' health-related absence rates. We then introduce our measure of health into a generalization of the canonical model of school value-added to estimate school effectiveness separately for healthy and unhealthy students. This paper then explores the characteristics that improve school effectiveness and investigates the policies that schools can implement to boost unhealthy students' academic outcomes.

There are several interesting findings in this paper. We find that value-added models that omit student health from their controls for student heterogeneity estimate a dispersion in school effectiveness that is biased upward by as much as 5%; while not very large, this bias is also nontrivial, indicating the importance of controlling for health heterogeneity when possible. We also find important heterogeneity in school value-added based on student health. The dispersion in school effectiveness is up to 31% larger among unhealthy students than among healthy students, indicating that schools are more influential in determining academic outcomes among unhealthy students. We also find that much of the across-school variation in the value-added gap is driven by variation in value-added among the unhealthy group, suggesting that school policy boosting effectiveness on unhealthy students can be an important way to close health-based achievement gaps.

Turning to a closer investigation of school policy, we find that while nurses have null effects on several student outcomes, homebound teachers - those that are qualified to provide at-home instruction to severely ill or disabled students - boost unhealthy students' test scores by 0.065 standard deviations. Paired with null effects on healthy students, our findings indicate that homebound teachers close the achievement gap between healthy and unhealthy students by 16%. Furthermore, for unhealthy students, homebound teachers decrease special education enrollment and time taken to graduate from high school, indicating success of the homebound education system at improving outcomes beyond just standardized exam performance. The substantial positive effects of homebound teachers that we document in this paper open the door for more school policy that is geared toward helping unhealthy students as well as more research to design and evaluate this policy.

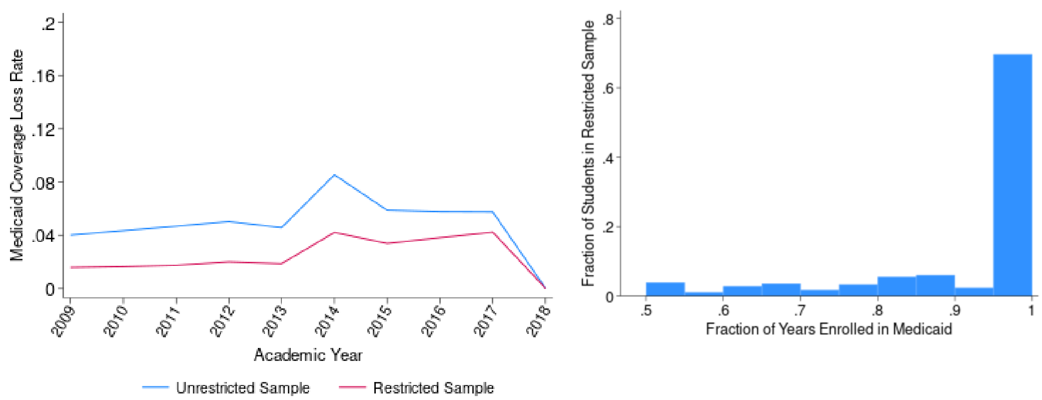
With a unique dataset that contains students' academic outcomes, demographic characteristics, and detailed health information, there are several fruitful avenues for future research. It would be interesting to estimate match effects between students and schools, separately by gender and separately by race. A large literature has shown that students of varying demographic characteristics are prone to different health conditions, and that the effects of poor health on educational and labor market outcomes also vary by demographics. The interaction between health and school inputs might also differ based on gender or race. Given the suggestive evidence in this paper that nurses may improve health, another interesting direction for future work is to identify whether schools generate causal gains to student health. There is much work to be done in understanding complementarities between health inputs and school inputs to the production of human capital.

1.A More on the Sample Restrictions

Decreasing the Medicaid Coverage Loss Rate

In this appendix, we first provide evidence that our sample restrictions decrease the Medicaid coverage loss rate among those ever observed to be enrolled in Medicaid. This is one of the main reasons for restricting the sample, since we can only observe student health during periods of Medicaid coverage. In Figure 1A.1, panel (a) shows that while Medicaid coverage loss was not high to begin with (4-8% of Medicaid-enrolled students lose coverage the following year), our sample restrictions decrease the coverage loss rate to 2-3%. Panel (b) shows that our sample restrictions - determined by Medicaid enrollment and subsidized lunch eligibility - result in a group of students almost always enrolled in Medicaid.

Figure 1A.1: Medicaid Coverage Loss by Year and Coverage Duration



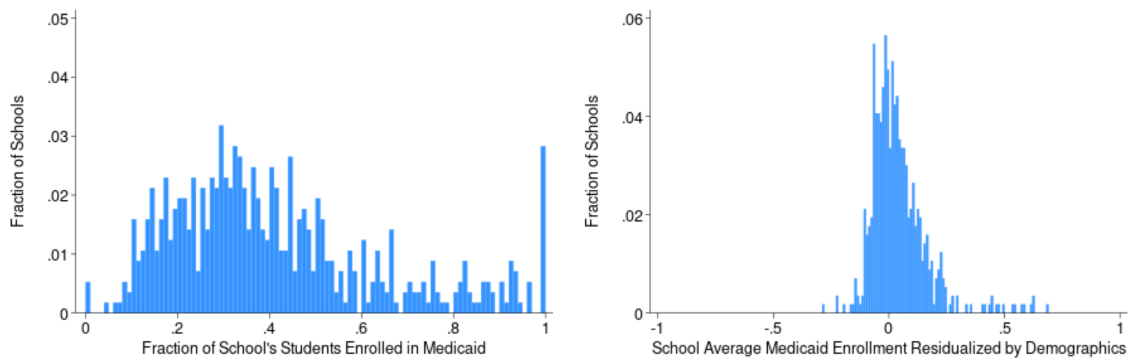
(a) Medicaid Coverage Loss Rate (b) Medicaid Enrollment Frequency

Notes: Panel (a) shows the Medicaid coverage loss rate, which is calculated in year t as the number of students that lost Medicaid coverage from t to $t + 1$ out of the total number of students enrolled in Medicaid in year t. Students that exit the school records are not considered to lose coverage, explaining the final year's zeros. The blue line contains all students that are observed at least once in the Medicaid data; the red line contains all students that are enrolled in Medicaid and eligible for subsidized lunch in at least half of the years they are observed. Panel (b) shows the fraction of years that students in the analysis sample are enrolled in Medicaid; by construction, this is always at least 0.5.

Potential Endogeneity of Selection into the Sample

We now address concern of endogeneity of selection into the sample that could bias our value-added estimates. One could imagine a scenario in which certain schools (or volunteers associated with the school) encourage parents to enroll their children in Medicaid, given the 2008 expansion that made all children eligible; this would mean selection into the sample could be endogenous to school enrollment. This could be a problem for identification of school effectiveness in the value-added model, especially if the students encouraged to enroll in Medicaid by their school end up with better health outcomes.[70] In Figure 1A.2, we show why this is not a concern. Panel (a) shows that in roughly 3% of high schools, every student (with an ACT score) is enrolled in Medicaid in 2018, which could mean that those schools do encourage Medicaid enrollment. However, panel (b) shows that once we residualize the Medicaid enrollment indicator by demographic characteristics (race/ethnicity, gender, subsidized lunch status, and indicators for English proficiency and special education enrollment), there is little unexplained variation. If there were schools close to one in panel (b), it would indicate students enrolling in Medicaid despite demographic characteristics from which we predict them not to enroll (e.g., they are from higher-income families). The fact that we do not see any schools remotely close to one (or minus one, which would suggest that some schools actually discourage Medicaid enrollment) is a strong sign that there is not systematic selection into our sample based on actions by schools.

Figure 1A.2: Histogram Over Fraction of Students Enrolled in Medicaid in 2018 (High Schools Only)



(a) Raw Medicaid Enrollment

(b) Medicaid Enrollment Residualized by Demographics

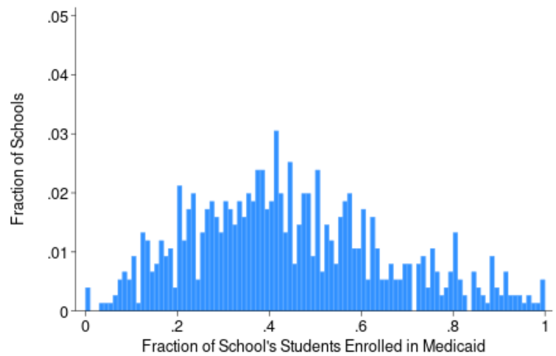
Notes: The sample includes all students with a 2018 grade 11 test score, since this is the high school value-added outcome. Panel (a) presents a histogram over the school-level fraction of students that are enrolled in Medicaid in 2018. The bar at the far right indicates that in about 3% of high schools, every student was enrolled in Medicaid at some point in 2018. We then residualize the student-level Medicaid enrollment indicator by the following demographic characteristics: race/ethnicity, gender, subsidized lunch status, and indicators for English proficiency and special education enrollment. Panel (b) shows a histogram over the school-level average residualized Medicaid enrollment. For a school very close to zero on the x-axis, demographic characteristics explain nearly all of the variation in Medicaid enrollment. For a school closer to the one, more students are enrolled in Medicaid when we would predict them not to enroll based on demographic characteristics alone.

In Figure 1A.3, we replicate the previous figure for elementary schools (top) and middle schools (bottom). We find similar patterns but with less variation than in high schools. The left panels on raw Medicaid enrollment indicate that there are very few elementary or middle schools with full Medicaid enrollment among their students, so the concern that elementary or middle schools encourage Medicaid enrollment is lower than in high schools. Once we residualize by demographic

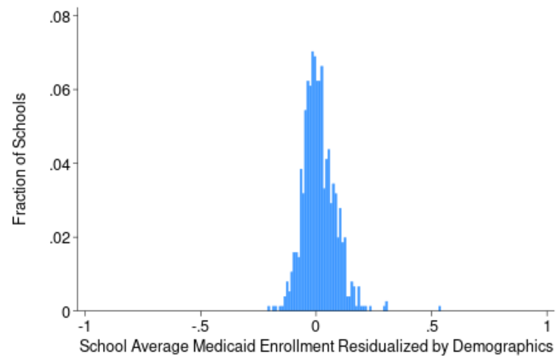
[70] In this scenario, schools that encouraged more Medicaid enrollment would systematically have healthier students, leading to conflation of the main value-added parameters with student health.

characteristics, even less variation remains here than in the high school case, indicating that the likelihood that schools play a role in Medicaid enrollment among younger students is extremely low. We find the same patterns in earlier years of the data, though we omit them for brevity.

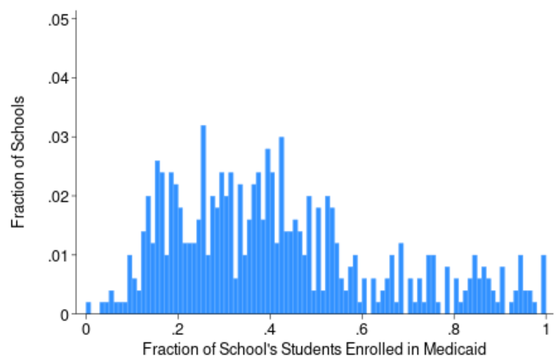
Figure 1A.3: Histogram over Fraction of Students Enrolled in Medicaid in 2018
(Top: Elementary Schools, Bottom: Middle Schools)



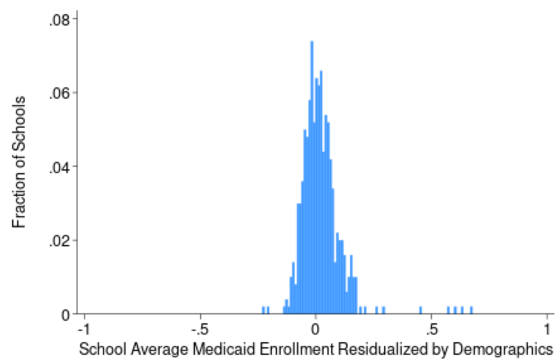
(a) Raw Medicaid Enrollment
(Elementary)



(b) Residualized
(Elementary)



(c) Raw Medicaid Enrollment
(Middle)



(d) Residualized
(Middle)

Notes: In panels (a) and (b), the sample includes all students with a 2018 grade 5 test score, since this is the elementary school value-added outcome. In panels (c) and (d), the sample includes all students with a 2018 grade 8 test score, since this is the middle school value-added outcome. Panels (a) and (c) present histograms over the school-level fraction of students that are enrolled in Medicaid in 2018. We then residualize the student-level Medicaid enrollment indicator by the following demographic characteristics: race/ethnicity, gender, subsidized lunch status, and indicators for English proficiency and special education enrollment. Panels (b) and (d) show histograms over the school-level average residualized Medicaid enrollment.

1.B Random Forest

Shortcomings of Alternative Methods to the Random Forest

One strand of the health literature (e.g., Adams et al., 2002; Tan et al., 2023) has used Johns Hopkins' ACG Grouper System to group people based on morbidity tied to the combination of diagnoses a person receives. Though a tried-and-tested method for understanding morbidity in adults (e.g., Weiner et al., 1991, 1992, 1996), Currie et al. (2010) find that the ACG Grouper System excludes several diagnoses that are highly prevalent among children, making this alternative method less appealing for our context. Another option would have been to use a health measure previously constructed in a related study; unfortunately, as Einav et al. (2022) put it, there is no well-established measure of health that we can simply take from the past literature. There is a strand of the health literature that has used random forests to study diagnosis codes associated with mental health problems (see Shatte et al., 2019, for a full review); though these studies generally focus on a much smaller subset of diagnosis codes than we do in this paper, they make a strong case for the generalizability of the random forest.

More Details on Constructing the Random Forest

We construct random forest models using absence rate as the outcome and all current- and previous-year diagnosis codes as predictors. This is done separately for elementary, middle, and high schools, because the production function of health and the way in which health translates into academic outcomes might differ substantially based on age. In building and testing a random forest, the first step is to separate the data into two pieces: the training set (to help the model learn) and the test set (to evaluate the model on an external piece of data); the reason for this is that typical models for prediction are only useful if they are externally valid. Due to a switch from ICD-9 to ICD-10 diagnosis codes in October 2015, we construct two separate models, one for each code type; as a result, we have two training sets and two test sets. For the ICD-9 model, we use 2008 - 2012 as the training set and 2013 - Sept. 2015 as the test set; for the ICD-10 model, we use Oct. 2015 - 2017 as the training set and 2018 as the test set.[71] This division ensures that we avoid evaluating predictions for students in later years using earlier data for the same students.

After separating the data into the training and test sets, the next step in the algorithm is a cross-validation procedure to optimize some of the model hyperparameters, which will limit over-fitting;[72] we optimize on number of decision trees in the random forest and maximum depth to which each tree can grow.[73] Because we are working with immense datasets (over forty thousand unique

[71]It is common to use either 67% or 80% of the data to form the training set; given that it is substantially more computationally costly to train than test, and taking into account the vast number of diagnosis codes we have as model predictors, we opt for smaller training sets to manage runtime.

[72]Over-fitting is an issue in which the random forest can only perform well on the training set but fails to make accurate predictions when faced with data beyond what it used to learn.

[73]In general, with more decision trees and a maximum depth closer to zero, random forests become more robust and encounter fewer issues with over-fitting.

diagnosis codes from over twenty five million Medicaid claims), it is computationally intractable to optimize the hyperparameters on the full training data, even after separating the ICD-9 and -10 codes from each other. To solve this issue, we use a 10% random sample of students from the ICD-9 (ICD-10) set to cross-validate for the ICD-9 (ICD-10) model. For the same reason of computational intractability, we are unable to train the random forest on the full data after the recovering the optimal hyperparameters from the cross-validation procedure. To solve this issue, we also use the cross-validation procedure to identify diagnosis codes that are most effective in separating the sample (known as feature importance); we take only the codes above mean importance. We implement 20 iterations of the following 3-fold cross-validation procedure outlined in Hastie et al. (2009) to identify the hyperparameters and important diagnosis codes:

1. Randomly select a number of trees $B \in [50, 500]$ and maximum depth $d \in [1, 100]$
2. Split the training set into 3 subsamples
3. For each subsample s, repeat the following:
    (a) Train a random forest on s with the hyperparameters selected in Step (1)
    (b) Test the model on the remaining data (training set without s)
4. Return the average performance of hyperparameters B and d
5. Return a vector of diagnosis codes above mean importance

We then take the highest-performing hyperparameters and the diagnosis codes above mean-importance to train the random forests (six separate models determined by two different diagnosis code types and three different school types) on the full data. Because we include previous-year diagnosis codes in the models, we are unable to recover predictions for two windows: the first year of the ICD-9 data and the first partial-year of the ICD-10 data (due to the mid-year code switch). We drop 2008 from our subsequent analyses of school value-added and school policy; it is only used for the purpose of previous-year diagnoses in the random forest and lagged variables in the value-added model. For missing absence rate predictions at the end of 2015 (due to the code switch), we carry forward previous-year predictions. From the random forest segment of our study, we recover health-related absence rate predictions for all students in our sample (those consistently enrolled in Medicaid and eligible for subsidized lunch) from 2009 through 2018, so this is our main analysis period for the estimation of school value-added.

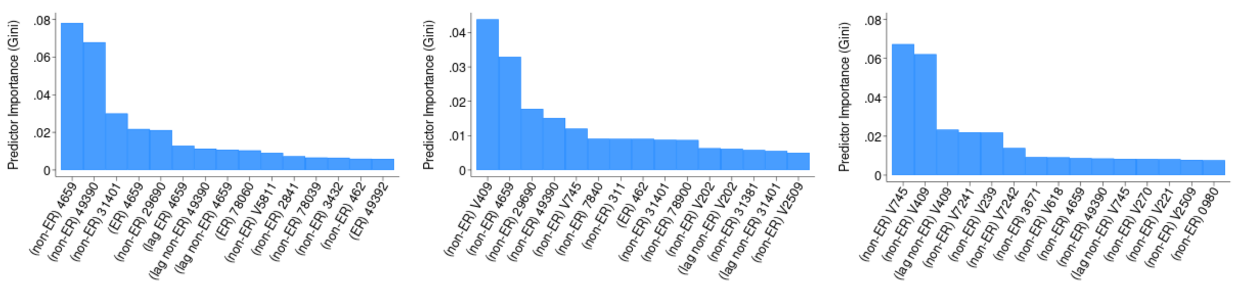
Table 1B.1: Results from the Cross-Validation Procedure

	Elementary school		Middle school		High school	
	ICD-9 (1)	ICD-10 (2)	ICD-9 (3)	ICD-10 (4)	ICD-9 (5)	ICD-10 (6)
	Cross-Validation Procedure					
Hyperparameters						
Number of trees	124	268	443	268	279	84
Maximum tree depth	14	24	19	44	20	20
Important diagnoses	881	1902	1178	2293	1558	2377
	Out-of-Sample Evaluation of the Final Models					
Mean squared error	0.003	0.004	0.006	0.007	0.020	0.025
R-squared	0.037	0.053	0.050	0.085	0.091	0.139

Notes: The top panel of this table reports the main results from the cross-validation procedure, including hyperparameters (number of decision trees in the random forest and maximum depth to which each tree can grow) and number of diagnosis codes above mean-importance for predicting students' absence rates. The bottom panel reports the mean squared error and R-squared of the final random forest models that use the optimized hyperparameters and diagnosis codes from the top panel.

More Intermediate Results from the Random Forest

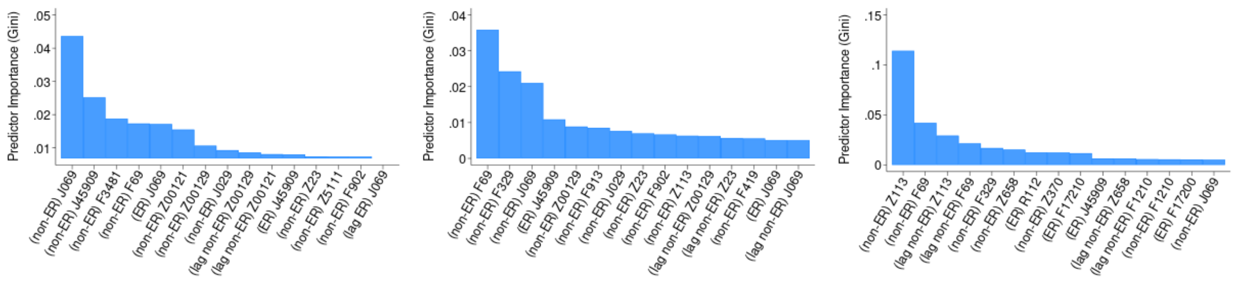
Figure 1B.1: Diagnosis Code Importance in the ICD-9 Models



(a) Elementary (b) Middle (c) High

Notes: These graphs show ICD-9 diagnosis codes that were most important in splitting the sample in the random forest algorithm. There are clear differences in the important diagnoses across school types, indicating that health varies substantially across students of varying age. Among elementary school students, we find 4659 (acute upper respiratory infection), 49390 (asthma), and 31401 (ADHD) to be most important in the model; among middle school students, we find V409 (disorder of personality and behavior), 4659 (acute upper respiratory infection), and 29690 (single episode mood disorder) to be most important; among high school students, we find V745 (screening for venereal disease) and V409 from current- as well as previous-year non-ER claims (disorder of personality and behavior) to be most important. And a quick note that these diagnosis codes are not necessarily the most common, just that they effectively split the sample at various levels of the tree.

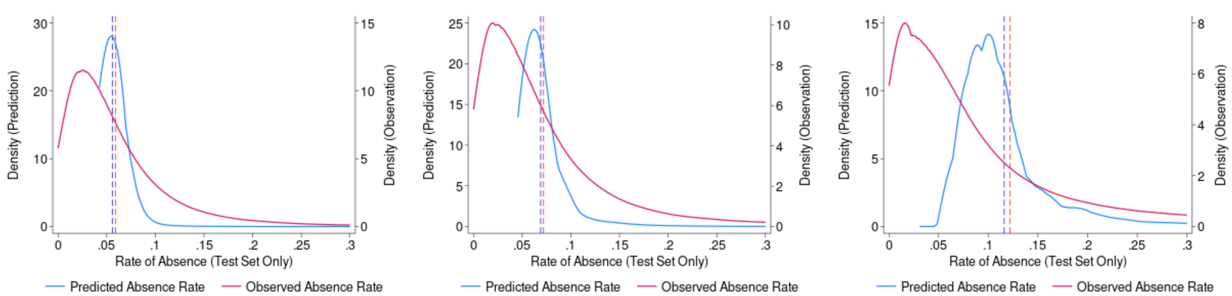
Figure 1B.2: Diagnosis Code Importance in the ICD-10 Models



(a) Elementary (b) Middle (c) High

Notes: These graphs show ICD-10 diagnosis codes that were most important in splitting the sample in the random forest algorithm. There are clear differences in the important diagnoses across school types, indicating that health varies substantially across students of varying age. Among elementary school students, we find J069 (acute upper respiratory infection), J45909 (asthma), and F3481 (disruptive mood dysregulation disorder) to be most important in the model; among middle school students, we find F69 (disorder of personality and behavior), F329 (single episode major depressive disorder), and J069 (acute upper respiratory infection) to be most important; among high school students, we find Z113 from current- as well as previous-year non-ER claims (screening for sexually transmitted infection) and F69 (disorder of personality and behavior) to be most important. And a quick note that these diagnosis codes are not necessarily the most common, just that they effectively split the sample at various levels of the tree.

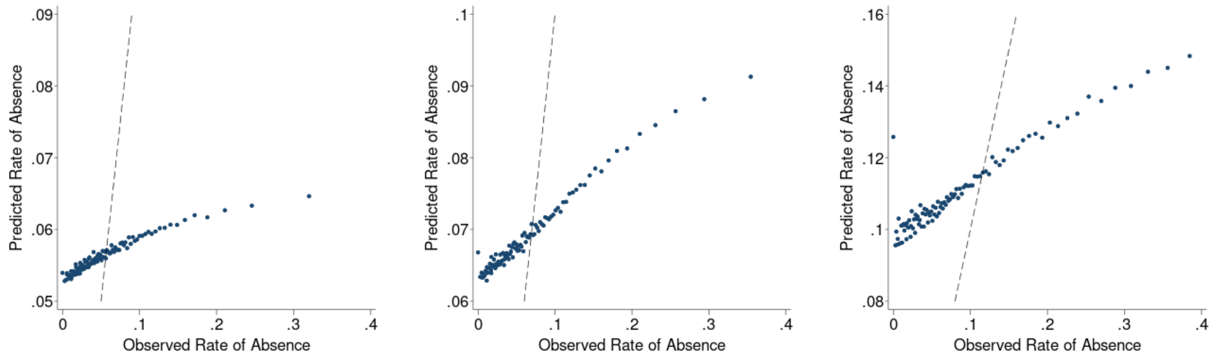
Figure 1B.3: (Out-of-Sample) Observed and Predicted Absence Rate PDFs



(a) Elementary (b) Middle (c) High

Notes: These graphs show out-of-sample predicted absence rate in blue and observed absence rate in red. The random forest shrinks the distribution toward the mean (which matches closely from observation to prediction, as depicted by the dashed vertical lines), likely due to factors that are predictive of absence but unobserved to the random forest. This is also guaranteed to happen to some extent, just based on the way the random forest recursively splits the sample and then predicts based on the mean of the final remaining subsample. We are not concerned with this, because our primary goal with the random forest is not to match the predictions to the data as closely as possible; rather, it is to construct a health index by which we can reliably separate healthy from unhealthy students.

Figure 1B.4: Binned Scatterplot of (Out-of-Sample) Predicted and Observed Absence Rates



(a) Elementary (b) Middle (c) High

Notes: These graphs show how predicted absence rates compare with observed absence rates using only observations in the test set. The flat slope (the 45-degree line is plotted as a dashed line) confirms that variation is lost when absence rates are pushed toward the mean, but the positive correlation indicates students with higher predicted health-related absence rates tend to have higher overall absence rates.

To show that our constructed health index contains more than what can be explained by student demographics, we estimate the following simple econometric model:

$$\hat{HI}_{it} = \beta_0 + X_{it}'\beta_1 + \beta_2\hat{HI}_{i,t-1} + \varepsilon_{it} \tag{1.21}$$

In Equation (1.21), we regress our constructed health index $\hat{HI}_{it}$ on a vector of student demographics X_{it} including race/ethnicity, sex, subsidized lunch status, and indicators for English proficiency and special education enrollment. We also include a lag on the health index, which will encompass what we think of as persistence in health conditions.

This regression shows us several important things. Demographic characteristics explain 89% of the variation in the health index among elementary school students, 80% of the variation among middle school students, and only 74% of the variation among high school students. This means that there is a substantial portion of variation in student health that is not captured by observables often present in education data, especially among older students. Substantial increases in R-squared when adding in a lag on the health index indicate the persistence of health over time; in other words, students that are healthier (unhealthier) tend to be healthier (unhealthier) in the following year. The lag on the health index also absorbs much of the observed association between demographics and health, though nearly all of the coefficients still remain statistically significant. Ultimately, the fact that there is substantial variation in predicted absence rates beyond what can be explained by observable demographic characteristics indicates the importance of our health index.

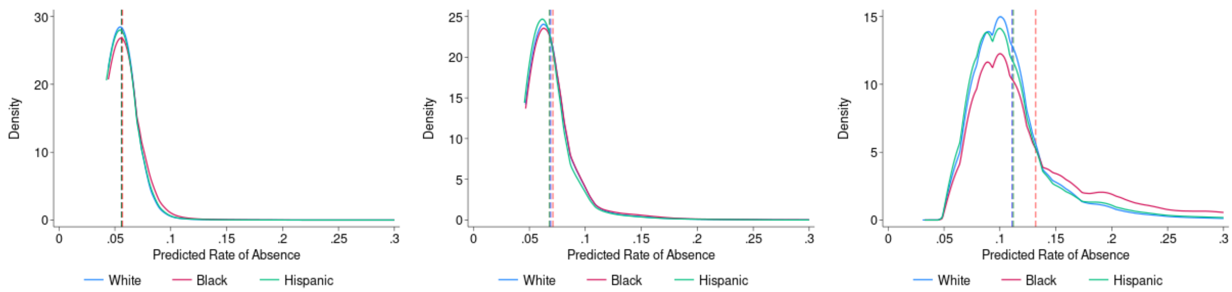
Table 1B.2: Explaining the Health Index with Demographic Characteristics

	Elementary school		Middle school		High school	
	(1)	(2)	(3)	(4)	(5)	(6)
Asian	0.290	0.056	0.313	0.022	1.368	0.376
	(0.010)	(0.007)	(0.023)	(0.019)	(0.038)	(0.027)
Black	0.629	0.180	0.897	0.303	3.985	1.485
	(0.005)	(0.004)	(0.011)	(0.010)	(0.019)	(0.014)
Hispanic	0.586	0.186	0.860	0.257	2.519	0.748
	(0.006)	(0.004)	(0.013)	(0.011)	(0.024)	(0.017)
Male	0.720	0.194	0.908	0.194	1.340	0.134
	(0.003)	(0.003)	(0.008)	(0.007)	(0.015)	(0.010)
Free lunch	4.915	1.291	5.891	1.891	8.437	2.256
	(0.003)	(0.004)	(0.008)	(0.010)	(0.015)	(0.012)
Reduced-price lunch	4.798	1.249	5.659	1.739	8.437	2.013
	(0.007)	(0.006)	(0.016)	(0.016)	(0.029)	(0.021)
Limited English proficient	−0.158	−0.134	−0.458	−0.203	−0.614	−0.013
	(0.007)	(0.005)	(0.018)	(0.015)	(0.037)	(0.026)
Special education	0.654	0.276	1.365	0.564	2.375	0.509
	(0.005)	(0.003)	(0.011)	(0.009)	(0.019)	(0.013)
Lagged health index		0.708		0.771		0.897
		(0.001)		(0.001)		(0.001)
Lagged health index		Y		Y		Y
R-squared	0.894	0.947	0.798	0.871	0.738	0.882
Observations	1,173,976	1,061,443	619,805	567,394	732,971	686,767

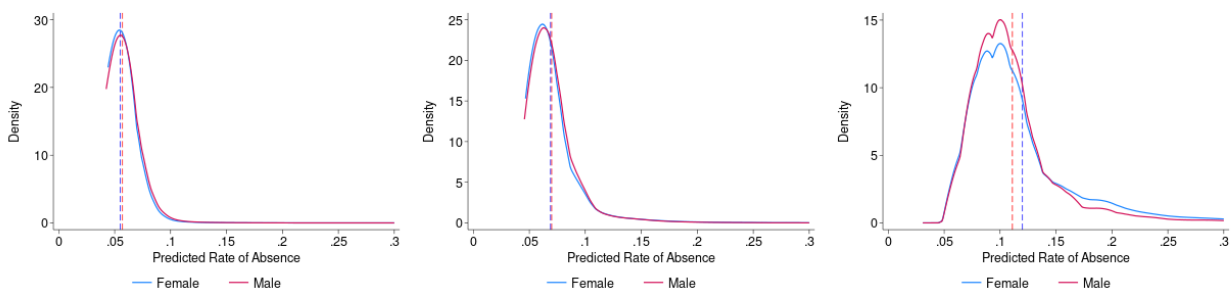
Notes: The predicted absence rate outcome and its lag have been multiplied by 100 for ease of interpretation; each coefficient represents a percentage point marginal change in predicted health-related absence rate. The sample included in each specification adheres to all of our primary restrictions: (1) students enrolled in Medicaid in at least half of their observations, (2) students eligible for subsidized lunch in at least half of their observations, (3) students for whom we observe test scores, and (4) students for whom we were able to predict an absence rate from their observed diagnosis codes. The specifications in columns (1), (3), and (5) omit the lagged health index while the specifications in column (2), (4), (6) include it. The R-squared in the specifications without the lagged health index indicate that demographic characteristics explain 89% of the variation in the health index among elementary school students, 80% of the variation among middle school students, and only 74% of the variation among high school students. Substantial increases in R-squared when adding in a lag on the health index indicate the persistence of health over time; in other words, students that are healthier (unhealthier) tend to be healthier (unhealthier) in the following year. The reference category on the race and ethnicity variable is White; the reference category on the subsidized lunch variable is ineligible (there are some of these still because we restrict to students eligible in at least half of their observations).

Figure 1B.5: Predicted Absence Rate by Demographic Subgroup

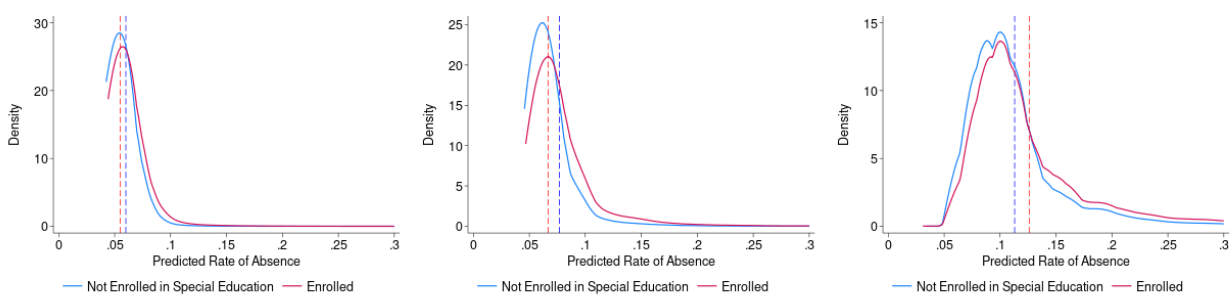
By race/ethnicity:



By gender:



By special education enrollment:



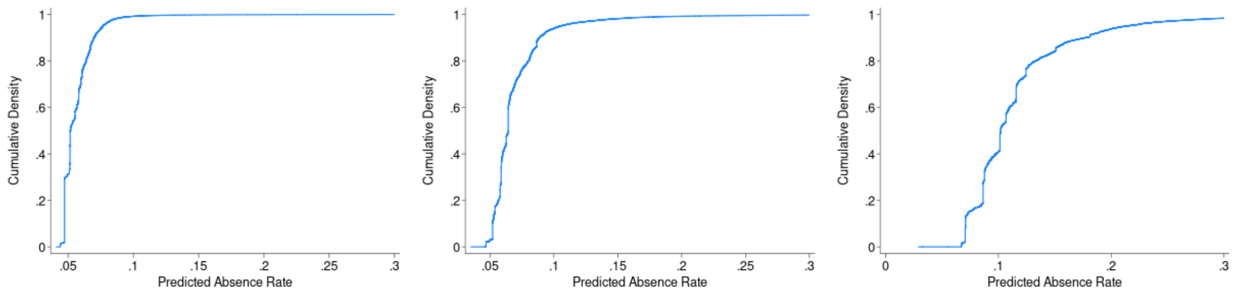
(a) Elementary (b) Middle (c) High

Notes: These graphs show distributions of predicted health-related absence rates by race/ethnicity (top), gender (middle), and special education enrollment (bottom). In general, the differences grow in high schools, indicating that more of the demographic heterogeneity across students in high schools is correlated with their health. There is still a large variation in health, even when conditioning on demographic characteristics of race/ethnicity or gender. In addition, the overall pattern holds even for the subgroups: there is a large concentration on the lower end of the distribution and a tail extending out to the right.

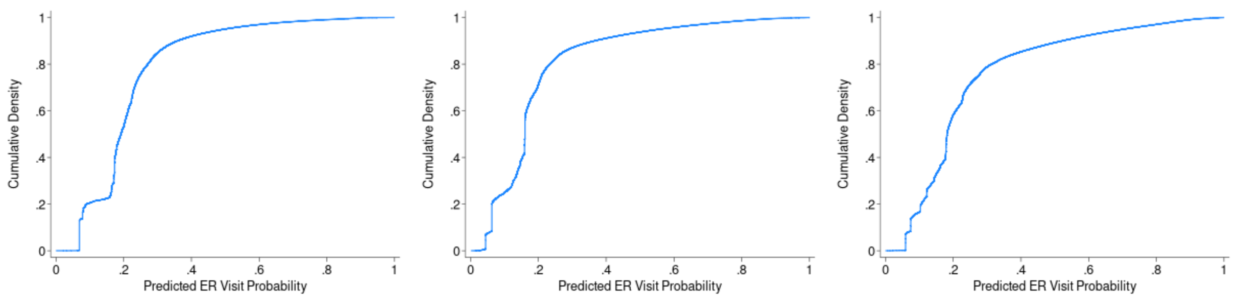
Robustness of the Random Forest to Other Outcomes

Figure 1B.6: Random Forest Prediction CDFs for Various Outcomes

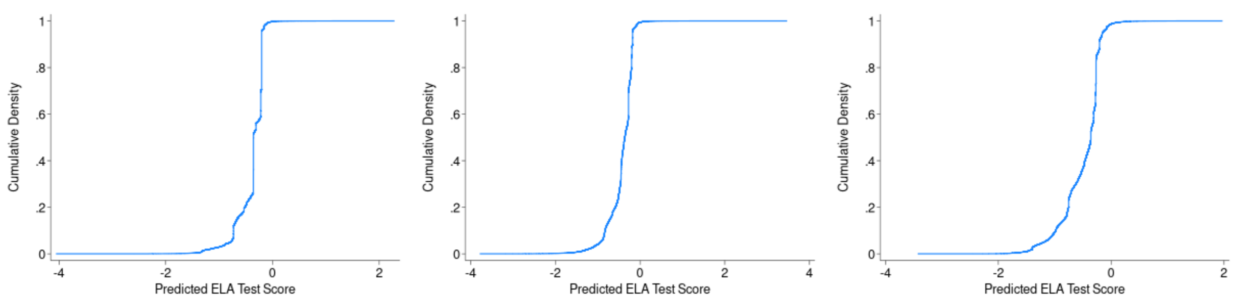
Using students' absence rates as the outcome:



Using an ER visit indicator as the outcome:



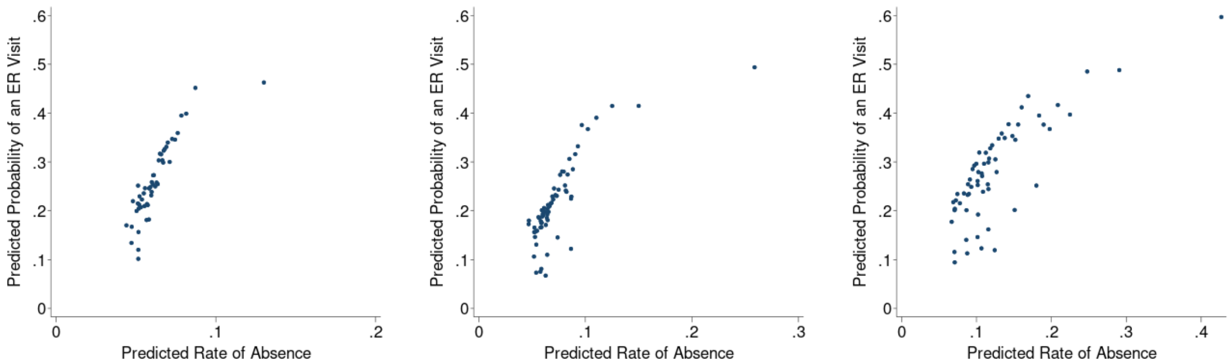
Using students' ELA test scores as the outcome:



(a) Elementary (b) Middle (c) High

Notes: We construct random forests to predict students' probabilities of ER visits and ELA test scores from their observed diagnosis codes, which serve as robustness checks to our baseline model that predicts students' absence rates. These graphs show CDFs of the random forest predictions of absence rates (top), ER visit probabilities (middle), and ELA test scores (bottom). Mirroring our main specification, we do this separately for students in elementary schools (left), middle schools (middle), and high schools (right). Using test scores as the outcome leads to the largest loss in observed variation, due to a number of factors beyond health that determine students' test scores.

Figure 1B.7: Correlation of Predicted Absence Rates with Predicted ER Visit Probabilities



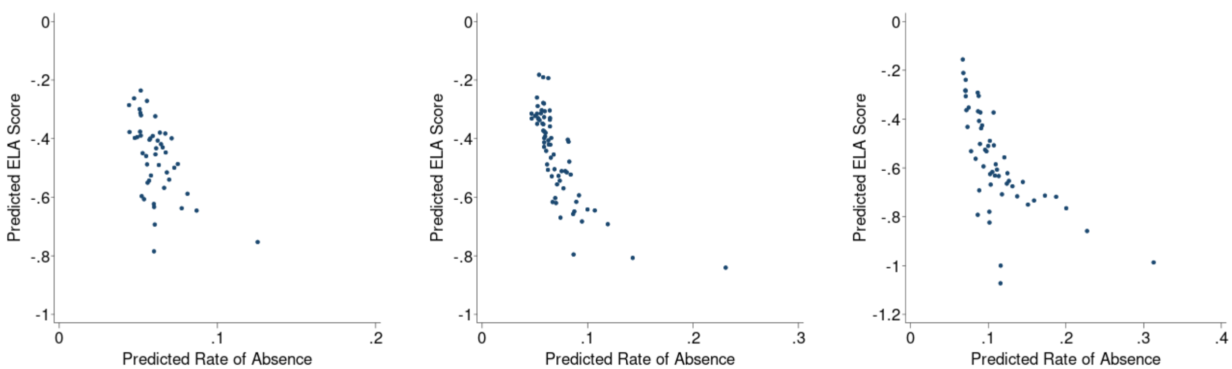
(a) Elementary
Correlation coef. = 0.478

(b) Middle
Correlation coef. = 0.416

(c) High
Correlation coef. = 0.444

Notes: These graphs depict binned scatterplots of random forest predictions of absence rates against predictions of emergency room visit probabilities. They show that the predicted absence rate predictions from our baseline specification of the random forest are correlated with predicted ER visits. In other words, students for whom our primary model predicts high absence rates also receive high ER visit probability predictions; these students are predicted to be less healthy in either model.

Figure 1B.8: Correlation of Predicted Absence Rates with Predicted ELA Test Scores



(a) Elementary
Correlation coef. = -0.381

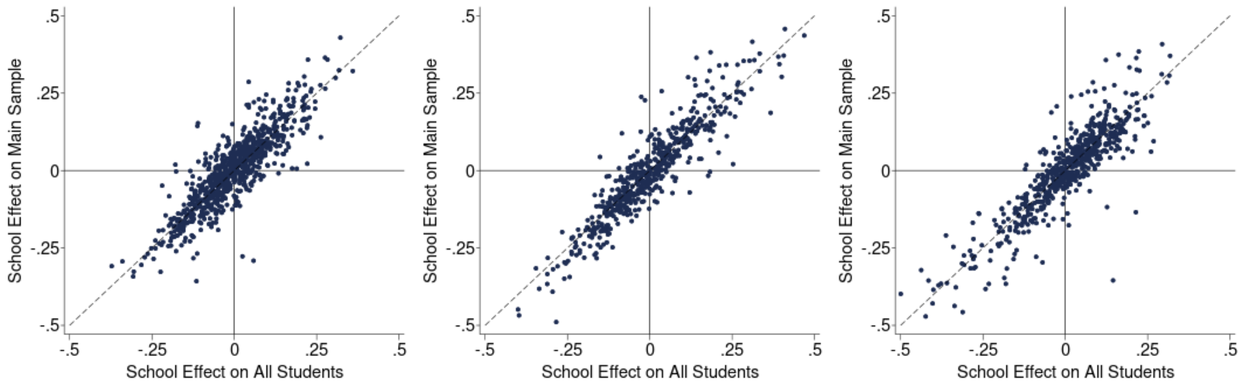
(b) Middle
Correlation coef. = -0.399

(c) High
Correlation coef. = -0.430

Notes: These graphs depict binned scatterplots of random forest predictions of absence rates against predictions of ELA test scores. They show that the predicted absence rate predictions from our baseline specification of the random forest are correlated with predicted test scores. In other words, students for whom our primary model predicts high absence rates also receive lower test score predictions; these students are predicted to be less healthy in either model.

1.C Supplementary Tables and Figures

Figure 1C.1: Correlation of Sample-Estimated School Effects with Population Estimates



(a) Elementary
Correlation coef. = 0.868

(b) Middle
Correlation coef. = 0.922

(c) High
Correlation coef. = 0.893

Notes: On the x-axis is the estimated school effects using all students in the school (population) and on the y-axis is the estimated school effect using students in our analysis sample, both using ELA test scores as the dependent variable in the value-added estimation in Equation (1.1). Each point represents a school. The 45-degree line is given as a reference to represent schools for which we estimate the same effect on students in our sample as on the population. The strong positive correlation indicates that selection into our sample does not drive our value-added results.

Table 1C.1: Dispersion in School Effects Estimated from our Sample and the Population

	Elementary school		Middle school		High school	
	Sample (1)	Population (2)	Sample (3)	Population (4)	Sample (5)	Population (6)
	Outcome: ELA test scores					
Dispersion	0.119	0.108	0.170	0.149	0.149	0.147
	(0.003)	(0.003)	(0.005)	(0.004)	(0.004)	(0.004)
Number of schools	828	854	558	576	606	638
	Outcome: Math test scores					
Dispersion	0.156	0.153	0.184	0.172	0.138	0.141
	(0.004)	(0.004)	(0.006)	(0.005)	(0.004)	(0.004)
Number of schools	828	854	558	576	606	638

Notes: This table reports dispersion of school effect estimates from the model of value-added in Equation (1.1). Estimates in Columns (1), (3), and (5) use only the students in our analysis sample while those in Columns (2), (4), and (6) use the population of Wisconsin public school students. Standard errors are calculated following Ahn and Fessler (2003). The small decrease in schools when estimating using the sample rather than the population indicates that the selection of schools into the distribution is minor; comparable estimates on the dispersion of school value-added alleviate any concern of the impact of selection of students into the sample on our results.

Table 1C.2: Coefficient Estimates for Key Variables from the Standard Value-Added Model (Equation (1.1)) and the Model that Controls for Health (Equation (1.2))

	Elementary school		Middle school		High school	
	Standard Model (1)	Health Included (2)	Standard Model (3)	Health Included (4)	Standard Model (5)	Health Included (6)
	Outcome: ELA test scores					
Past Math score	0.479	0.474	0.197	0.191	0.258	0.251
	(0.013)	(0.013)	(0.007)	(0.007)	(0.006)	(0.006)
Past ELA score	1.302	1.300	0.657	0.650	0.650	0.644
	(0.013)	(0.013)	(0.007)	(0.007)	(0.006)	(0.006)
Black	0.033	0.014	0.054	0.057	−0.118	−0.125
	(0.114)	(0.114)	(0.082)	(0.082)	(0.058)	(0.058)
Male	−0.035	−0.036	−0.188	−0.184	−0.055	−0.064
	(0.048)	(0.048)	(0.037)	(0.037)	(0.036)	(0.036)
Health index	-	−0.011	-	−0.009	-	−0.011
		(0.004)		(0.002)		(0.001)
Academic year fixed effect	Y	Y	Y	Y	Y	Y
R-squared	0.694	0.699	0.680	0.700	0.736	0.744
Observations	131,552	131,392	140,921	140,781	142,699	142,501
	Outcome: Math test scores					
Past Math score	1.341	1.331	0.641	0.633	0.727	0.720
	(0.014)	(0.014)	(0.007)	(0.007)	(0.032)	(0.006)
Past ELA score	0.350	0.348	0.188	0.183	0.181	0.179
	(0.013)	(0.013)	(0.007)	(0.007)	(0.005)	(0.029)
Black	0.207	0.167	−0.147	−0.127	−0.206	−0.217
	(0.121)	(0.121)	(0.089)	(0.089)	(0.054)	(0.054)
Male	0.074	−0.009	0.093	0.104	0.038	0.029
	(0.051)	(0.013)	(0.040)	(0.040)	(0.033)	(0.033)
Health index	-	−0.014	-	−0.019	-	−0.009
		(0.004)		(0.002)		(0.001)
Academic year fixed effect	Y	Y	Y	Y	Y	Y
R-squared	0.664	0.672	0.658	0.669	0.757	0.763
Observations	131,486	131,325	140,763	140,624	143,139	142,940

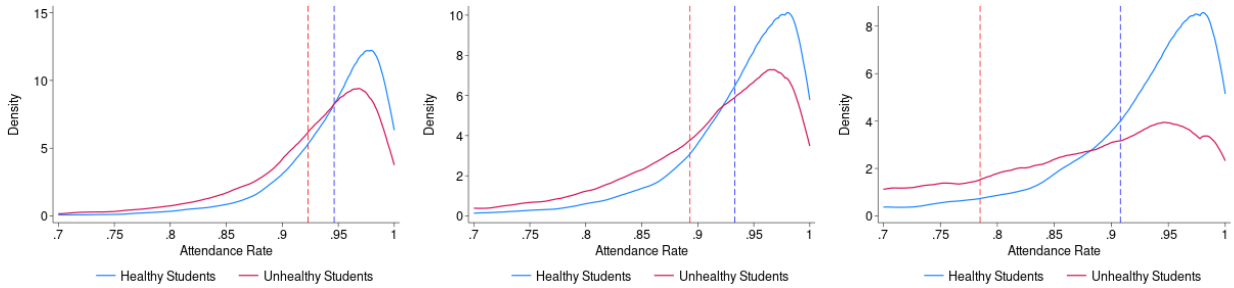
Notes: This table reports some of the main coefficient estimates on covariates in the X_{it} and H_{it} vectors from Equations (1.1) and (1.2), which control for student heterogeneity. Columns (1), (3), and (5) present estimates from the standard model of value-added; Columns (2), (4), and (6) present estimates from the model that controls for student health.

Table 1C.3: Health Group Membership by Year, 2009-2018

Academic Year	Total Students (1)	Percent Healthy (2)	Percent Unhealthy (3)
		Elementary school	
2009	18,683	71.2	28.5
2010	20,127	70.2	29.5
2011	20,149	69.1	30.7
2012	20,666	67.7	32.0
2013	21,073	68.7	31.1
2014	22,298	64.7	35.0
2015	23,098	76.4	23.4
2016	24,201	62.2	37.6
2017	25,549	61.8	37.7
2018	25,122	61.7	36.3
		Middle school	
2009	17,286	60.0	39.7
2010	17,339	62.1	37.6
2011	18,779	62.2	37.6
2012	18,792	60.0	39.7
2013	20,231	60.6	39.1
2014	20,165	56.3	43.4
2015	20,600	71.3	28.5
2016	20,817	78.2	21.5
2017	22,249	77.2	22.5
2018	23,050	74.4	24.1
		High school	
2009	16,839	77.6	21.9
2010	16,450	76.2	23.9
2011	17,429	78.2	21.6
2012	17,285	77.5	22.2
2013	18,857	77.5	22.2
2014	18,738	56.2	43.5
2015	17,684	68.2	31.5
2016	19,720	54.0	45.8
2017	19,648	53.4	46.2
2018	19,987	54.6	44.2

Notes: This table includes student-year observations with an outcome used in the value-added estimation (grade 5, grade 8, and grade 11 students). Columns (2) and (3) do not add up to 100 due to a small fraction of students for which we are unable to construct a health index.

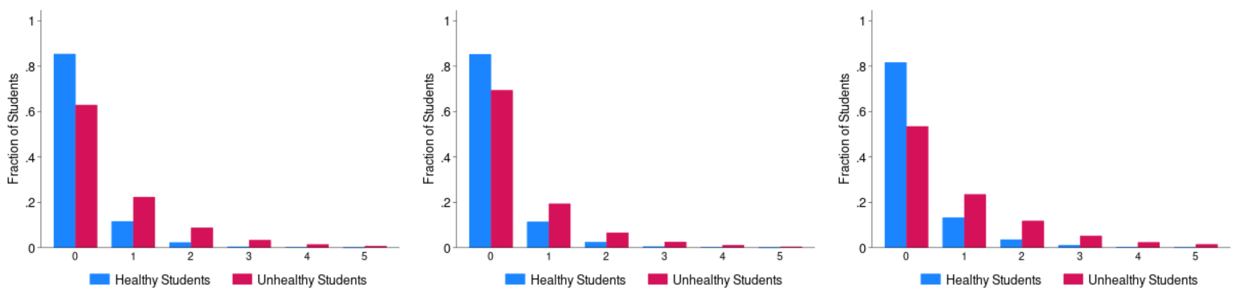
Figure 1C.2: Observed Attendance Rate by Health Group (2009 - A Training Set Year)



(a) Elementary (b) Middle (c) High

Notes: These graphs present observed attendance rate distributions in 2009 among students in each health group. The blue line contains healthy students while the red line contains unhealthy students.

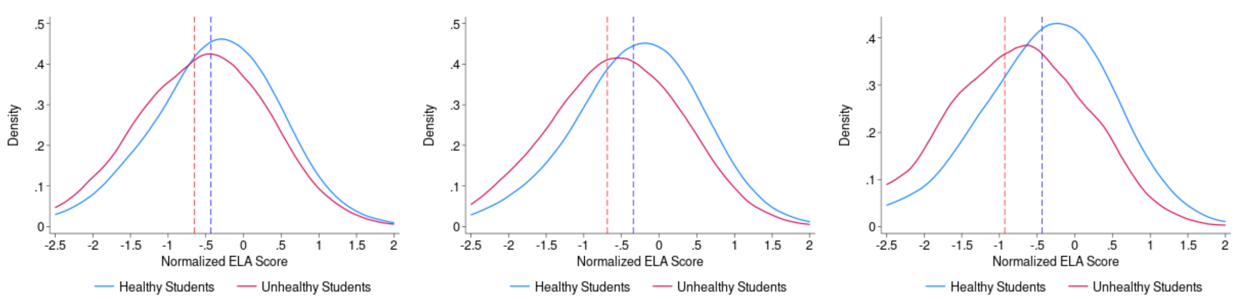
Figure 1C.3: Emergency Department Visits by Health Group (2009)



(a) Elementary (b) Middle (c) High

Notes: These graphs present the frequency with which students from different health groups visited the emergency department in 2009. The blue bars contain healthy students while the red bars contain unhealthy students. We truncate at 5, but some students visit the emergency department more frequently.

Figure 1C.4: Normalized ELA Test Scores by Health Group (2009)



(a) Elementary (b) Middle (c) High

Notes: These graphs present ELA test score distributions in 2009 among students in each health group. The blue line contains healthy students while the red line contains unhealthy students.

Table 1C.4: Coefficient Estimates for Key Variables from the Main Value-Added Specification (Equation (1.4))

	Elementary school		Middle school		High school	
	ELA	Math	ELA	Math	ELA	Math
	(1)	(2)	(3)	(4)	(5)	(6)
Past Math score	0.604	1.504	0.180	0.539	−0.071	0.631
	(0.119)	(0.127)	(0.063)	(0.068)	(0.032)	(0.030)
Past ELA score	1.356	0.190	0.674	0.256	1.037	0.205
	(0.121)	(0.129)	(0.063)	(0.068)	(0.031)	(0.029)
Cumulative years eligible	−0.007	0.011	−0.012	0.008	−0.007	−0.028
for subsidized lunch	(0.007)	(0.008)	(0.006)	(0.006)	(0.005)	(0.004)
Black	−0.014	0.087	0.008	−0.135	−0.167	−0.248
	(0.103)	(0.428)	(0.080)	(0.087)	(0.056)	(0.051)
Hispanic	0.067	−0.052	0.037	0.037	0.105	−0.021
	(0.096)	(0.615)	(0.083)	(0.090)	(0.069)	(0.064)
Male	−0.016	0.092	−0.202	0.095	−0.050	0.017
	(0.043)	(0.047)	(0.036)	(0.039)	(0.034)	(0.032)
Special education	−0.237	−0.085	−0.327	−0.416	−0.159	−0.314
	(0.080)	(0.086)	(0.067)	(0.072)	(0.060)	(0.055)
Health index	−0.020	−0.033	−0.022	−0.051	−0.038	−0.059
	(0.010)	(0.011)	(0.008)	(0.009)	(0.003)	(0.003)
Number of ER visits	−0.015	−0.006	−0.023	−0.010	−0.015	−0.005
	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
Academic year fixed effect	Y	Y	Y	Y	Y	Y
R-squared	0.673	0.634	0.680	0.649	0.732	0.754
Observations	140,618	140,535	144,470	144,320	148,573	149,011

Notes: This table reports some of the main coefficient estimates on covariates in the X_{it} and H_{it} vectors from Equation (1.4), which control for student heterogeneity. Columns (1), (3), and (5) present estimates from regressions that use English Language Arts (ELA) test scores as the dependent variable; Columns (2), (4), and (6) use Math test scores as the dependent variable. Confirming past literature, we find past test scores to be the most important covariate; we also find our constructed health index (the predicted health-related absence rate) to be strongly (and statistically significantly) predictive of a decrease in test scores. The increase in R-squared from elementary to middle to high school may suggest that additional non-cognitive inputs not included in the model are more important for standardized test-taking in earlier years.

Table 1C.5: Means and Standard Deviations of Covariates in Equation (1.15)

	Elementary school	Middle school	High school
	(1)	(2)	(3)
Employee counts			
School nurses	0.185	0.211	0.251
	(0.195)	(0.317)	(0.913)
Homebound teachers	0.009	0.005	0.019
	(0.111)	(0.018)	(0.155)
Psychologists	0.278	0.267	0.224
	(0.292)	(0.335)	(0.556)
Teacher demographics			
Percent White	0.958	0.929	0.946
	(0.084)	(0.139)	(0.120)
Percent Black	0.013	0.042	0.030
	(0.046)	(0.121)	(0.101)
Percent female	0.861	0.708	0.565
	(0.060)	(0.118)	(0.147)
Teacher qualifications			
Graduate degree	0.482	0.441	0.459
	(0.156)	(0.176)	(0.205)
Experience	14.2	13.7	13.6
	(2.4)	(3.0)	(3.8)
Student demographics			
Percent White	0.616	0.607	0.657
	(0.266)	(0.321)	(0.308)
Percent Black	0.116	0.173	0.158
	(0.183)	(0.291)	(0.273)
Percent female	0.490	0.491	0.497
	(0.066)	(0.087)	(0.137)
Percent special ed.	0.193	0.214	0.209
	(0.077)	(0.101)	(0.124)
Percent unhealthy	0.305	0.323	0.349
	(0.068)	(0.086)	(0.164)
Number of Schools	762	477	517
Number of Students	155,003	165,044	181,651

Notes: This table reports means and standard deviations of key variables used in Equation (1.15). This is crucial because we use standard deviation units of each of these variables in our descriptive exercise to understand how school effectiveness is influenced by a one standard deviation increase in each variable. These numbers differ slightly from Table 1.1 containing descriptive statistics, because this table restricts to students for which we are able to estimate a health index.

Table 1C.6: Coefficient Estimates for Variables in Equation (1.15), Healthy Subgroup

	Elementary school		Middle school		High school	
	ELA	Math	ELA	Math	ELA	Math
	(1)	(2)	(3)	(4)	(5)	(6)
Employee counts						
School nurses	−0.010	−0.008	0.011	0.013	−0.034	−0.048
	(0.004)	(0.005)	(0.006)	(0.007)	(0.013)	(0.014)
Homebound teachers	0.051	−0.022	−0.006	−0.012	0.009	0.012
	(0.013)	(0.018)	(0.005)	(0.006)	(0.006)	(0.006)
Psychologists	0.0001	−0.004	−0.012	−0.017	0.014	−0.001
	(0.004)	(0.006)	(0.006)	(0.007)	(0.006)	(0.006)
Teacher demographics						
Percent White	0.003	−0.004	−0.074	−0.041	−0.002	−0.009
	(0.006)	(0.008)	(0.017)	(0.020)	(0.013)	(0.013)
Percent Black	−0.008	−0.018	−0.023	0.0003	−0.004	−0.007
	(0.007)	(0.010)	(0.021)	(0.024)	(0.015)	(0.015)
Percent female	0.007	0.015	0.037	0.012	−0.003	0.003
	(0.004)	(0.006)	(0.009)	(0.010)	(0.008)	(0.008)
Teacher qualifications						
Graduate degree	−0.001	−0.005	−0.005	−0.002	−0.002	0.009
	(0.004)	(0.006)	(0.008)	(0.009)	(0.006)	(0.006)
Experience	−0.001	0.009	0.032	0.031	−0.015	0.001
	(0.004)	(0.006)	(0.008)	(0.010)	(0.008)	(0.008)
Student demographics						
Percent White	0.011	0.028	0.048	0.048	−0.004	−0.004
	(0.007)	(0.010)	(0.014)	(0.016)	(0.009)	(0.009)
Percent Black	−0.007	0.004	−0.016	−0.031	−0.011	−0.017
	(0.008)	(0.011)	(0.019)	(0.022)	(0.012)	(0.012)
Percent female	−0.005	−0.003	−0.005	−0.007	0.016	−0.014
	(0.007)	(0.009)	(0.016)	(0.012)	(0.011)	(0.012)
Percent special ed.	−0.012	−0.018	−0.024	−0.007	0.006	0.017
	(0.006)	(0.008)	(0.012)	(0.014)	(0.010)	(0.010)
Percent unhealthy	−0.009	0.002	−0.0003	−0.007	−0.074	−0.056
	(0.005)	(0.007)	(0.010)	(0.012)	(0.009)	(0.009)
R-squared	0.149	0.121	0.221	0.211	0.229	0.256
Number of schools	762	762	477	477	517	517

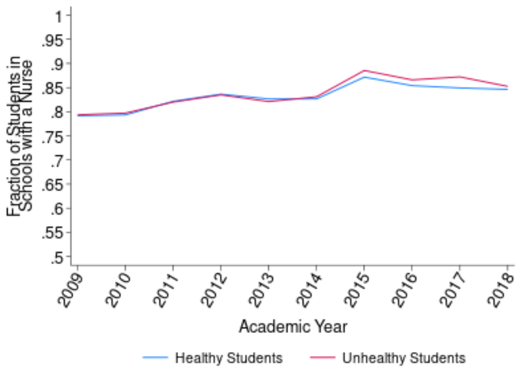
Notes: This table reports coefficient estimates from Equation (1.15) using school effectiveness among *healthy* students as the outcome. Columns (1), (3), and (5) present estimates from regressions that use ELA test scores as the dependent variable; Columns (2), (4), and (6) use Math test scores as the dependent variable. Standard errors are in parenthesis. The R-squared values do not match those in Table 1.10 because those come from bootstrapped values.

Table 1C.7: Coefficient Estimates for Variables in Equation (1.15), Unhealthy Subgroup

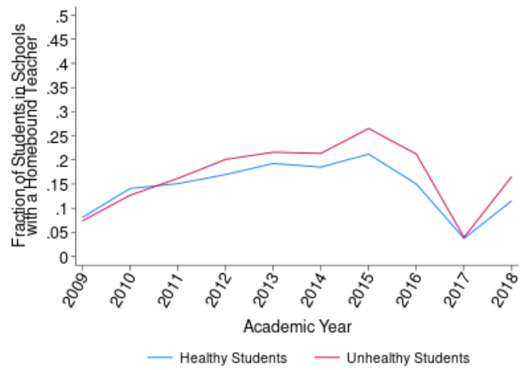
	Elementary school		Middle school		High school	
	ELA (1)	Math (2)	ELA (3)	Math (4)	ELA (5)	Math (6)
Employee counts						
School nurses	−0.006	−0.008	0.013	0.009	−0.028	−0.015
	(0.005)	(0.006)	(0.006)	(0.007)	(0.017)	(0.015)
Homebound teachers	0.029	−0.025	0.004	−0.008	−0.011	0.001
	(0.019)	(0.024)	(0.005)	(0.005)	(0.006)	(0.006)
Psychologists	0.004	−0.005	−0.009	−0.015	0.014	−0.002
	(0.005)	(0.006)	(0.006)	(0.007)	(0.007)	(0.004)
Teacher demographics						
Percent White	0.002	−0.001	−0.047	−0.040	−0.001	0.012
	(0.008)	(0.009)	(0.017)	(0.018)	(0.014)	(0.012)
Percent Black	−0.001	−0.009	−0.023	−0.023	−0.017	−0.004
	(0.008)	(0.010)	(0.020)	(0.021)	(0.014)	(0.013)
Percent female	0.016	0.013	0.041	0.022	−0.001	−0.0003
	(0.006)	(0.007)	(0.009)	(0.009)	(0.009)	(0.009)
Teacher qualifications						
Graduate degree	−0.003	−0.003	−0.015	−0.007	−0.010	−0.0005
	(0.006)	(0.007)	(0.008)	(0.009)	(0.008)	(0.007)
Experience	0.005	0.014	0.035	0.031	0.007	0.009
	(0.006)	(0.007)	(0.009)	(0.009)	(0.009)	(0.009)
Student demographics						
Percent White	0.015	0.014	0.045	0.062	−0.008	−0.003
	(0.009)	(0.011)	(0.015)	(0.016)	(0.011)	(0.010)
Percent Black	−0.010	−0.013	−0.002	−0.003	−0.005	−0.013
	(0.010)	(0.012)	(0.019)	(0.020)	(0.013)	(0.012)
Percent female	−0.003	−0.005	−0.008	−0.021	0.021	0.017
	(0.009)	(0.011)	(0.016)	(0.017)	(0.013)	(0.012)
Percent special ed.	−0.021	−0.023	−0.018	−0.022	0.009	0.004
	(0.008)	(0.009)	(0.013)	(0.014)	(0.011)	(0.010)
Percent unhealthy	−0.013	−0.001	−0.012	−0.019	−0.054	−0.032
	(0.006)	(0.008)	(0.010)	(0.011)	(0.007)	(0.007)
R-squared	0.111	0.122	0.180	0.257	0.294	0.294
Number of schools	762	762	477	477	517	517

Notes: This table reports coefficient estimates from Equation (1.15) using school effectiveness among *unhealthy* students as the outcome. Columns (1), (3), and (5) present estimates from regressions that use ELA test scores as the dependent variable; Columns (2), (4), and (6) use Math test scores as the dependent variable. Standard errors are in parenthesis. The R-squared values do not match those in Table 1.10 because those come from bootstrapped values.

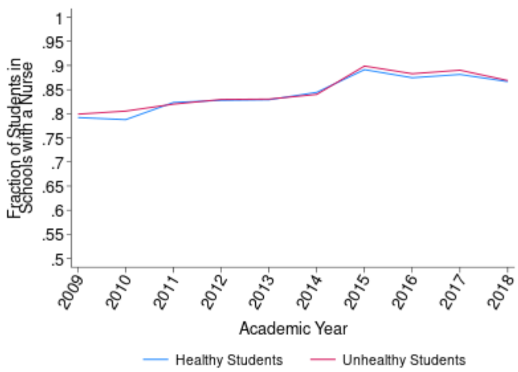
Figure 1C.5: Variation in Nurses (Left) and Homebound Teachers (Right) Across Time



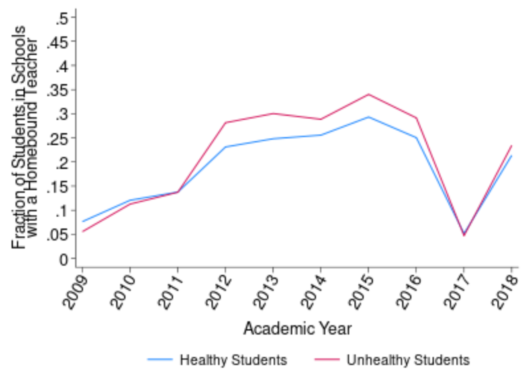
(a) Nurses in Elementary Schools



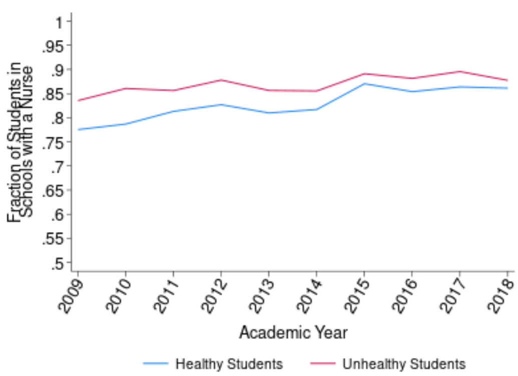
(b) Homebound Teachers in Elementary Schools



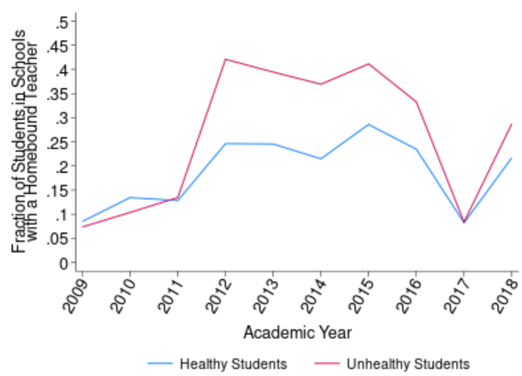
(c) Nurses in Middle Schools



(d) Homebound Teachers in Middle Schools



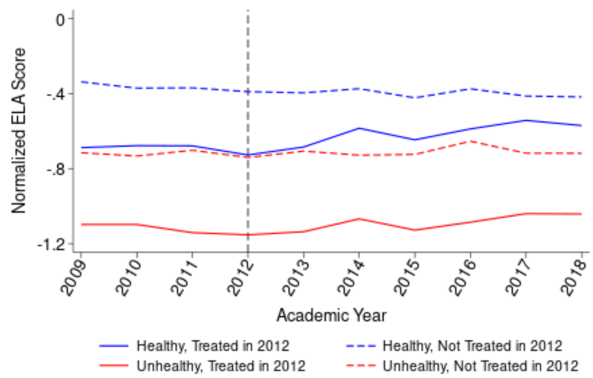
(e) Nurses in High Schools



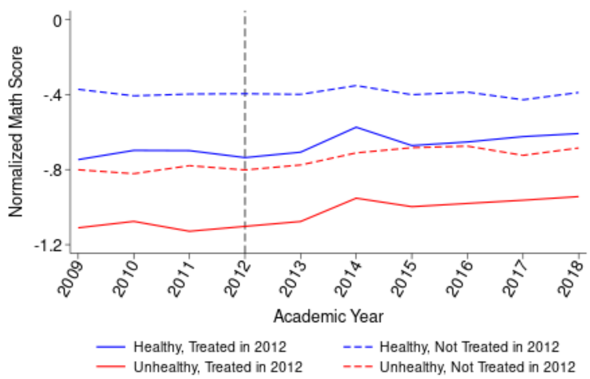
(f) Homebound Teachers in High Schools

Notes: Each graph shows trends in the fraction of healthy (blue) and unhealthy (red) students that attend schools with at least one nurse (left) or homebound teacher (right). The consistent pattern that more unhealthy students have access to nurses and homebound teachers suggests that the decision to hire these employees may be tied to the percentage of a school's students that are unhealthy, making this an important factor to control for in the two-way fixed effects model.

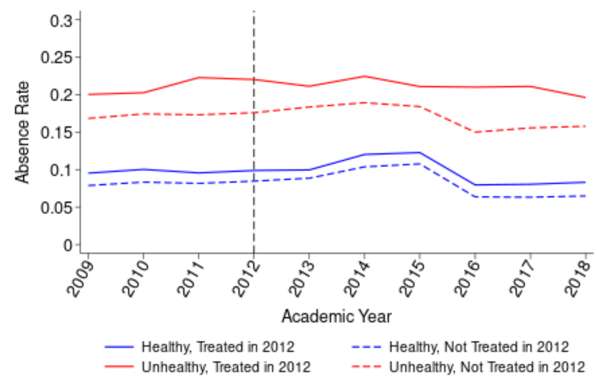
Figure 1C.6: Parallel Trends in Student Outcomes (Homebound Teacher Policy)



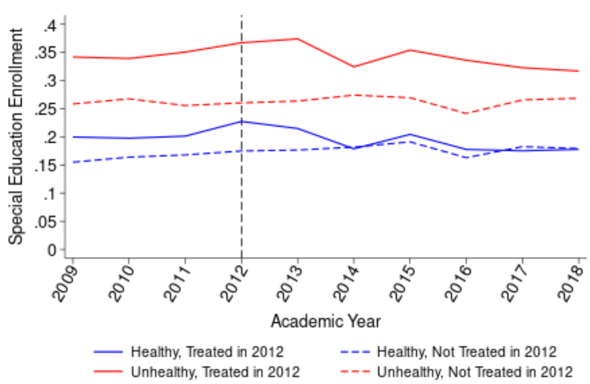
(a) ELA Test Scores



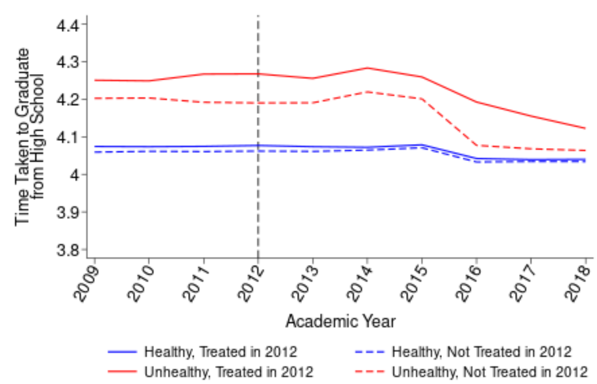
(b) Math Test Scores



(c) Absence Rates



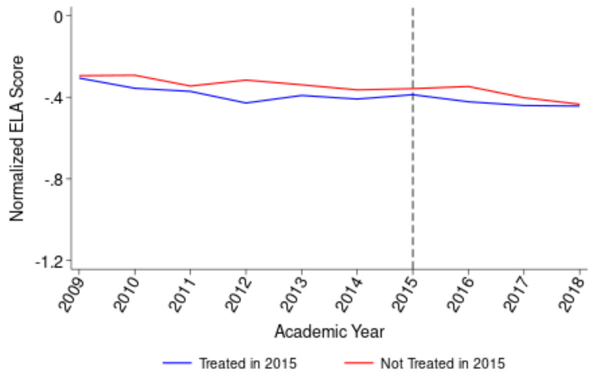
(d) Special Education Enrollment



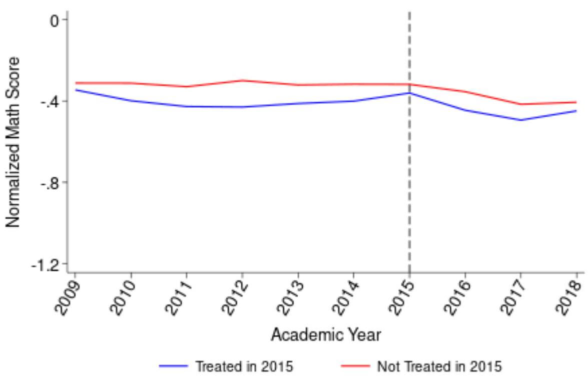
(e) Time Taken to Graduate from High School

Notes: Each graph shows outcome trends among high school students; schools that hired homebound teachers prior to 2012 are already treated and not included. The solid blue line is for healthy students in schools with a new homebound teacher in 2012; the dotted blue line is for healthy students in schools without a new homebound teacher in 2012; the solid red line is for unhealthy students in schools with a new homebound teacher in 2012; the dotted red line is for unhealthy students in schools without a new homebound teacher in 2012. For parallel trends, we compare the solid and dashed lines of a single color. Test score trends presented in the main text are replicated here for completeness. 2012 is presented because we find the largest increase in homebound teacher employment during this year, leading to larger sample sizes contained in each line for treated subgroups.

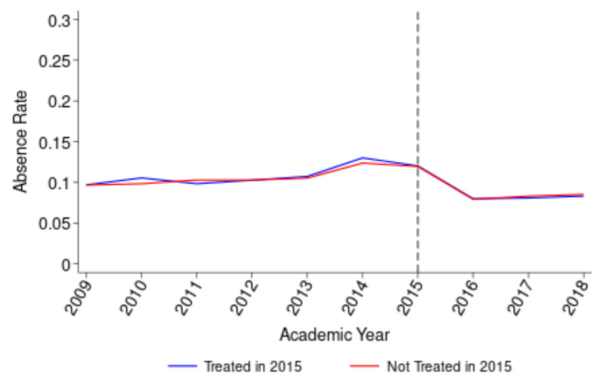
Figure 1C.7: Parallel Trends in Student Outcomes (School Nurse Policy)



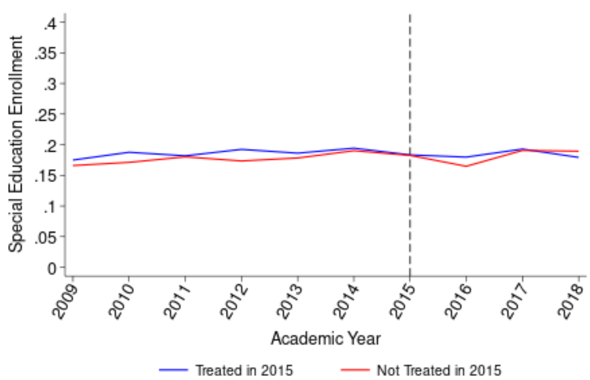
(a) ELA Test Scores



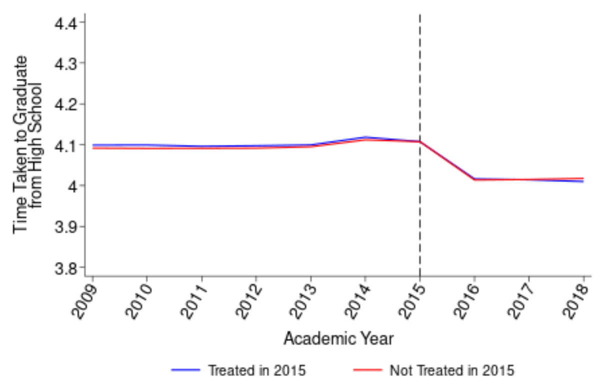
(b) Math Test Scores



(c) Absence Rates



(d) Special Education Enrollment



(e) Time Taken to Graduate from High School

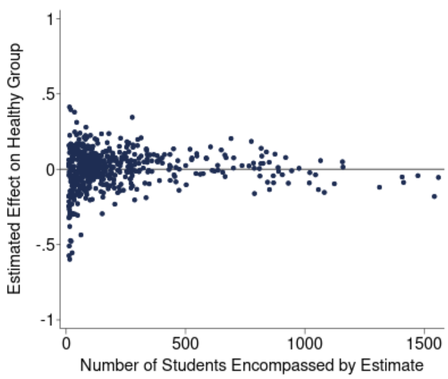
Notes: Each graph shows outcome trends among high school students; schools that hired nurses prior to 2015 are already treated and not included. The solid blue line is for students in schools with a new nurse in 2015; the solid red line is for students in schools with a new nurse in 2015. We do not break down subgroups further by health status, because nurses likely affect health group placement over time. 2015 is presented because we find the largest increase in nurse employment during this year, which leads to larger sample sizes in each line for treated subgroups.

1.D Robustness to Sampling Error and Shrinkage

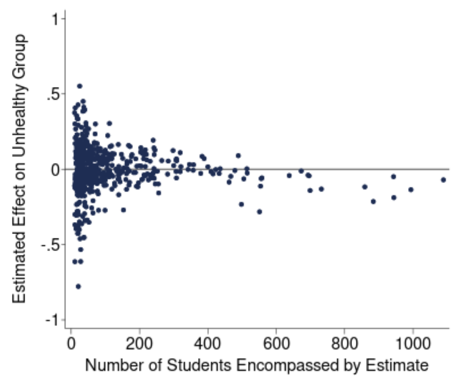
The Empirical Bayes Shrinkage Procedure

Under the selection-on-observables assumption, we recover unbiased estimates of θ_{jh}, the subgroup-specific school effects. However, these estimates are noisy, especially when there are very few students accounted for in the estimate (Kane and Staiger, 2002). In Figure 1D.1, we replicate a piece of the analysis from Aaronson et al. (2007), plotting the estimated subgroup-specific school effects against the number of students in the school-subgroup cell.

Figure 1D.1: Dependence of Fixed Effect Estimates on Cell Size



(a) School Effects on Healthy Group



(b) School Effects on Unhealthy Group

Notes: these graphs present fixed effect estimates of high schools, using Math test scores as the outcome. The dispersion in fixed effect estimates among school-subgroups of the same size is decreasing sharply in number of students in the school-subgroup. In other words, estimates that account for a smaller school-subgroup cell size tend to be noisier.

Unsurprisingly, there is substantially more variation in school effects for lower school-subgroup enrollment counts; the concern is that a few good or bad draws have the potential to heavily influence the estimated school effect when the school-subgroup cell size is small, leading to overall inflation of the dispersion in school effects. It is common in past literature to adjust for sampling error with Empirical Bayes (EB) strategies that shrink the estimates toward a common Bayesian prior (which is typically the distribution mean).

In models that use school value-added as a right-hand side variable of a regression to understand how school effectiveness impacts on future outcomes (for example, earnings in Chetty et al., 2014b), using the unbiased but noisy estimates creates attenuation bias toward zero as a result of classical measurement error. Using estimates that are purged of sampling error corrects attenuation bias by introducing non-classical measurement error (Angrist et al., 2023). In models that use school value-added on the left-hand side to understand what drives school effectiveness, the unbiased estimates are preferred; Hausman (2001) finds that biased estimates on the left-hand side can lead to biased and inconsistent coefficient estimates.

We only use models of the latter type in our analyses, placing value-added estimates on the left-hand side of regressions to understand how school characteristics and policy drive school effectiveness. For this reason, we do not require an Empirical Bayes procedure for any of our subsequent analyses. However, we include it here to show that the patterns we find in value-added dispersion are not driven by sampling error. Since the number of healthy and unhealthy students is different within each school, and by construction of the health groups the healthy group is on average about twice as large as the unhealthy group, the estimated school effects on unhealthy students will be noisier. We will show that differences in dispersion across health groups are not driven by this differential sampling error by also presenting results of the shrunk estimates.

Adapting the EB analysis from Bacher-Hicks and Koedel (2023) to our context, we "shrink" the school value-added estimates separately by health group; because the mean of the distribution changes with the shrinkage procedure due to asymmetric sampling error, we shrink the θ_{jh} estimates that have not yet been de-meaned. Each fixed effect estimate is shrunk by taking a weighted average of the estimate and the distribution mean:

$$\theta^*_{jh} = a_{jh}\hat{\theta}_{jh} + (1 - a_{jh})\bar{\theta}_h \tag{1.22}$$

The weight a_j is estimated as follows:

$$\hat{a}_{jh} = \frac{\hat{\sigma}^2_h}{\hat{\sigma}^2_h + \hat{\lambda}^2_{jh}} \tag{1.23}$$

$\hat{\sigma}^2_h$ is an estimate of the variance of value-added across all schools for health group h (for this we use the variance of estimated subgroup-specific school fixed effects)[74] while $\hat{\lambda}^2_{jh}$ is an estimate of the variance of the error in school j's value-added estimate (for this we use the squared standard error in the school-specific estimate). Estimates are noisier when λ_{jh} is larger and will be shrunk further toward the mean. As mentioned, the unhealthy group in a school is on average about half the size of the healthy group, so we expect school effects on the unhealthy group (particularly in the tails of the distribution) to be shrunk more forcefully toward the mean. If our results on the dispersion of school effectiveness hold after the shrinkage procedure, then we can conclude that the results are not driven by sampling error at the ends of the distribution (this is a sufficient condition, but not a necessary condition).

We then de-mean the shrunk estimates separately by subgroup such that each distribution is mean zero. We first take the mean of the shrunk estimates, separately by health group:

$$\bar{\theta}^*_h = \frac{1}{J}\sum_j \left[\theta^*_{jh}\right] \quad \text{for } h = 0, 1 \tag{1.24}$$

[74]One could also follow Angrist et al. (2023) and use the error-corrected variance of value-added across schools, which subtracts out the sampling variance to correct for the inflation in variance that it generates. In either case, we would recover an estimate of variance in value-added across schools that is biased (though in the latter case it is consistent). We go the simpler route here in estimating the variance of value-added because our main purpose is simply to show that patterns across schools are not driven by sampling error. We leave the EB section in the appendix rather than moving it up to the main text since most of our work focuses on individual schools.

Each shrunk estimate is then shifted by the relevant mean. E^*_{jh} represents the error-adjusted effect of school j on students in health group h:

$$E^*_{jh} = \theta^*_{jh} - \bar{\theta}^*_h \quad \text{for } j = 1, ..., J \text{ and } h = 0, 1 \tag{1.25}$$

Results Using Shrunk School Effect Estimates

Table 1D.1: Dispersion in School Effects Across Models (Robustness to Table 1.3)
(No Interaction Between the School Fixed Effect and Health Group Fixed Effect)

	Elementary school		Middle school		High school	
	Standard Model (1)	Health Included (2)	Standard Model (3)	Health Included (4)	Standard Model (5)	Health Included (6)
	Outcome: ELA test scores					
Non-shrunk ($\hat{\theta}_{j\hat{h}}$)	0.119	0.117	0.170	0.164	0.149	0.144
	(0.003)	(0.003)	(0.005)	(0.005)	(0.004)	(0.004)
EB shrunk ($\hat{\theta}^*_{j\hat{h}}$)	0.088	0.074	0.133	0.126	0.108	0.098
	(0.002)	(0.002)	(0.004)	(0.004)	(0.003)	(0.003)
Number of schools	828	828	558	558	606	606
	Outcome: Math test scores					
Non-shrunk ($\hat{\theta}_{j\hat{h}}$)	0.156	0.154	0.184	0.179	0.138	0.132
	(0.004)	(0.004)	(0.006)	(0.005)	(0.004)	(0.004)
EB shrunk ($\hat{\theta}^*_{j\hat{h}}$)	0.126	0.112	0.148	0.142	0.100	0.092
	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
Number of schools	828	828	558	558	606	606

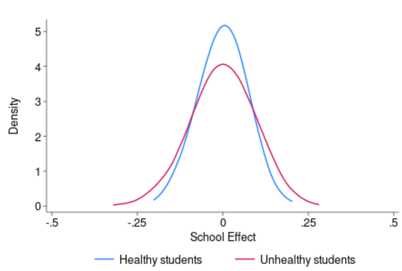
Notes: This table reports dispersion of school effect estimates from two models of value-added. Here we include non-shrunk and Empirical Bayes shrunk estimates. Estimates in Columns (1), (3), and (5) come from the standard model in Equation (1.1) while those in Columns (2), (4), and (6) come from the model that controls for health, described in Equation (1.2). Standard errors are calculated following Ahn and Fessler (2003), though one could also bootstrap the standard errors. Differences in dispersion of school effectiveness across the two models become more pronounced when health is included in the value-added estimation.

Table 1D.2: Dispersion in Estimated School Effects (Robustness to Table 1.6) (Main Interacted Specification in Equation (1.4))

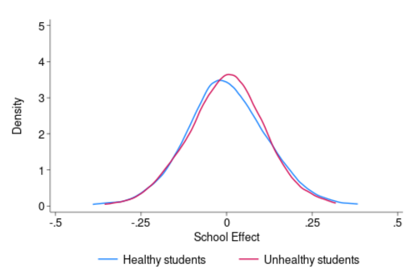
	Elementary school		Middle school		High school	
	Healthy students (1)	Unhealthy students (2)	Healthy students (3)	Unhealthy students (4)	Healthy students (5)	Unhealthy students (6)
	Outcome: ELA test scores					
Non-shrunk ($\hat{\theta}_{j\hat{h}}$)	0.115	0.151	0.148	0.157	0.136	0.163
	(0.003)	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)
EB shrunk ($\hat{\theta}^*_{j\hat{h}}$)	0.066	0.090	0.114	0.109	0.085	0.107
	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)	(0.003)
Number of schools	764	764	496	496	525	525
	Outcome: Math test scores					
Non-shrunk ($\hat{\theta}_{j\hat{h}}$)	0.152	0.178	0.176	0.175	0.119	0.147
	(0.004)	(0.005)	(0.006)	(0.006)	(0.004)	(0.005)
EB shrunk ($\hat{\theta}^*_{j\hat{h}}$)	0.103	0.118	0.133	0.123	0.081	0.096
	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
Number of schools	763	763	497	497	525	525

Notes: This table reports dispersion of non-shrunk and Empirical Bayes shrunk subgroup-specific school effect estimates. The non-shrunk estimates are what was presented in Table 1.6, while the shrunk estimates serve as robustness. Columns (1), (3), and (5) are for healthy students; Columns (2), (4), and (6) are for unhealthy students. We only include schools for which we estimate effects on *both* subgroups (healthy and unhealthy). Standard errors are calculated following Ahn and Fessler (2003), which describe standard errors on estimators of the standard deviation of a normal distribution; this is appropriate in our context because we make normality assumptions in the Empirical Bayes shrinkage procedure, though one could also bootstrap the standard errors.

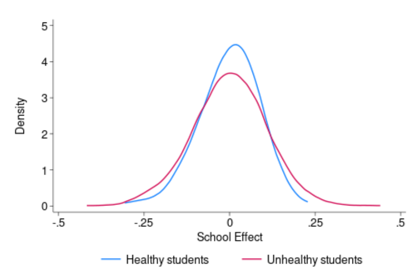
Figure 1D.2: Distributions of Shrunk School Effect Estimates, by Health Group



(a) Elementary



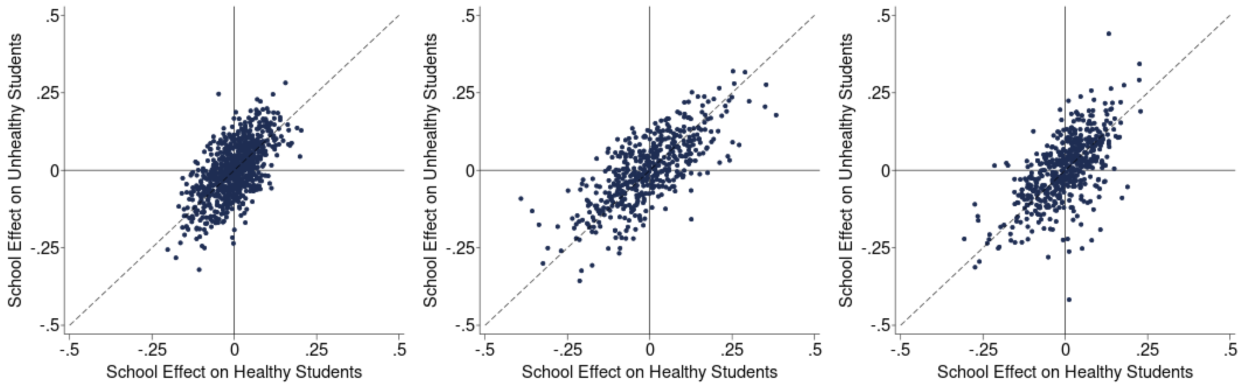
(b) Middle



(c) High

Notes: Each graph shows the health-based distribution of de-meaned and shrunk school effect estimates. For brevity we only include distributions that use ELA test scores as the outcome, though we find similar patterns in the regressions that use Math as the outcome are similar.

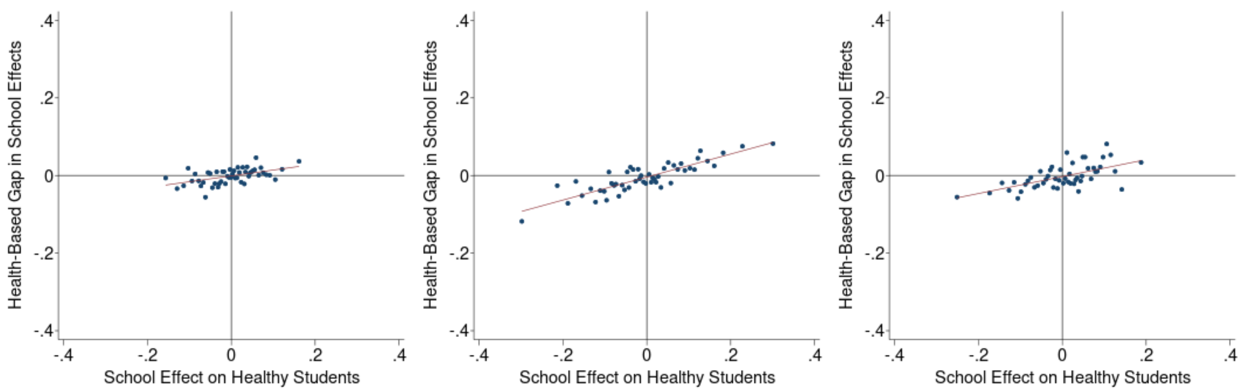
Figure 1D.3: Correlation of the Shrunk School Effects Across Health Groups



(a) Elementary (b) Middle (c) High

Notes: On the x-axis is the shrunk estimated school effect on healthy students and on the y-axis is the shrunk estimated school effect on unhealthy students, both using ELA test scores as the dependent variable in the value-added estimation. Each point represents a school. The 45-degree line is given as a reference to represent schools that have an equal effect on healthy and unhealthy students.

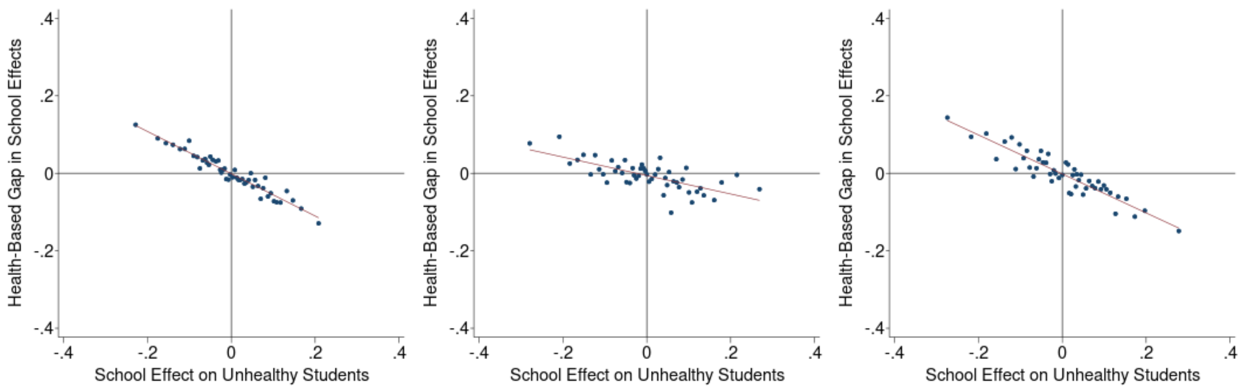
Figure 1D.4: Explaining the Gap in School Effectiveness (Shrunk): Healthy Students



(a) Elementary (b) Middle (c) High

Notes: On the x-axis is the shrunk school effect on healthy students using ELA test scores as the dependent variable, and on the y-axis is the value-added gap. This shows additional robustness of the weak relationship between the gap and the effect on healthy students.

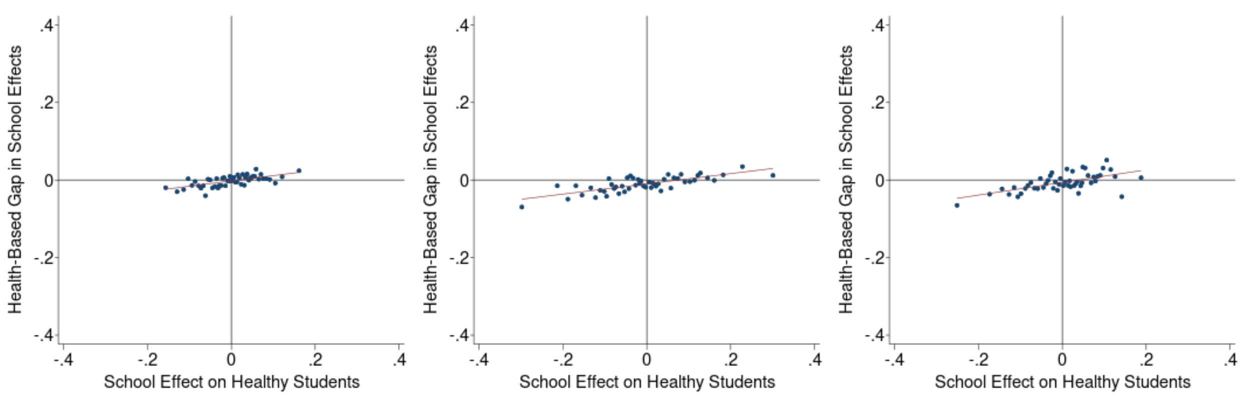
Figure 1D.5: Explaining the Gap in School Effectiveness (Shrunk): Unhealthy Students



(a) Elementary (b) Middle (c) High

Notes: On the x-axis is the shrunk school effect on unhealthy students using ELA test scores as the dependent variable, and on the y-axis is the value-added gap. This shows additional robustness of the stronger relationship between the gap and the effect on unhealthy students.

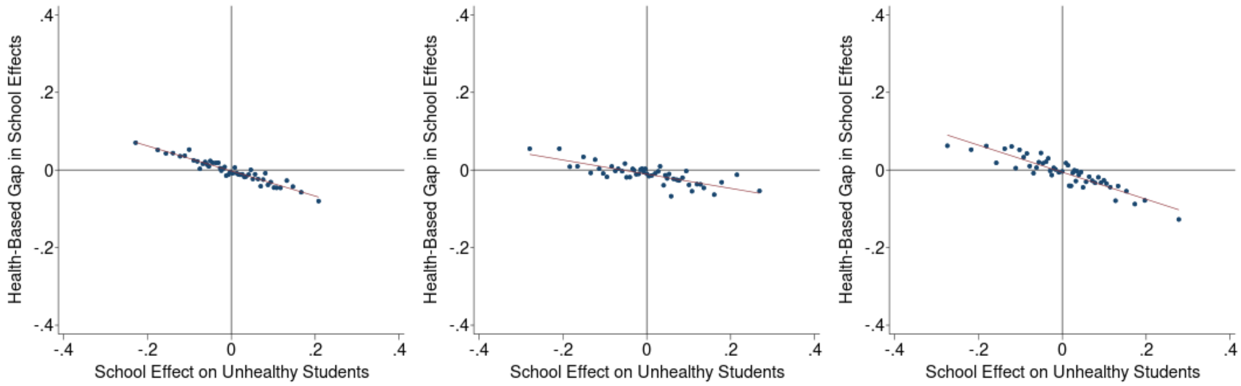
Figure 1D.6: Shrinking the Gap Directly
Explaining this Shrunk Gap in School Effectiveness: Healthy Students



(a) Elementary (b) Middle (c) High

Notes: On the x-axis is the shrunk school effect on healthy students using ELA test scores as the dependent variable, and on the y-axis is the value-added gap. Here, we estimate the gap as the difference in school effects across health groups prior to shrinkage, then we shrink the gap directly. This shows additional robustness of the weak relationship between the gap and the effect on healthy students.

Figure 1D.7: Shrinking the Gap Directly
Explaining this Shrunk Gap in School Effectiveness: Unhealthy Students



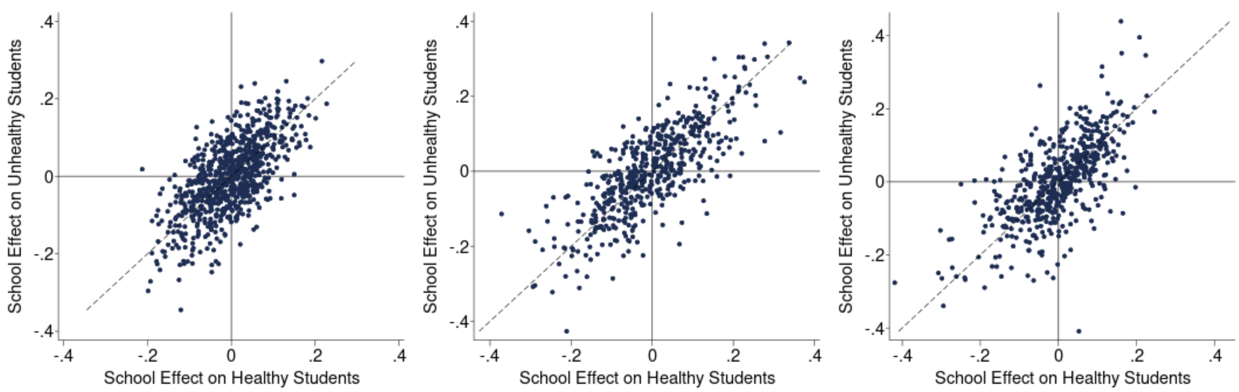
(a) Elementary (b) Middle (c) High

Notes: On the x-axis is the shrunk school effect on unhealthy students using ELA test scores as the dependent variable, and on the y-axis is the value-added gap. Here, we estimate the gap as the difference in school effects across health groups prior to shrinkage, then we shrink the gap directly. This shows additional robustness of the stronger relationship between the gap and the effect on unhealthy students.

1.E Robustness of the Sample to Relaxed Restrictions

Here, we include students enrolled in Medicaid anytime and eligible for subsidized lunch anytime, rather than students consistently enrolled in Medicaid and eligible for subsidized lunch.

Figure 1E.1: Correlation of the School Effects Across Health Groups - Broader Sample



(a) Elementary (b) Middle (c) High

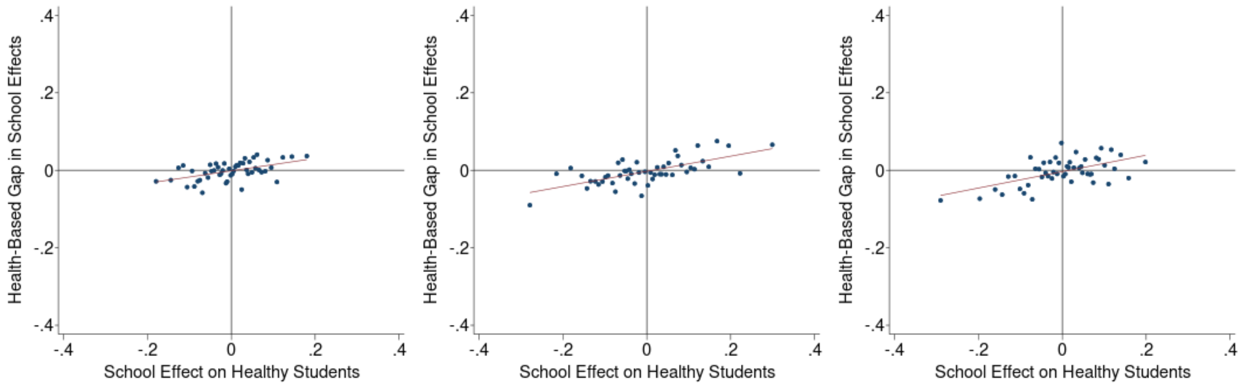
Notes: On the x-axis is the estimated school effect on healthy students and on the y-axis is the estimated school effect on unhealthy students, both using ELA test scores as the dependent variable in the value-added estimation. Each point represents a school. The 45-degree line is given as a reference to represent schools that have an equal effect on healthy and unhealthy students. We use the broader sample of students anytime enrolled in Medicaid and anytime eligible for subsidized lunch (as opposed to half-time for each); similar results indicate robustness of the sample.

Table 1E.1: Dispersion in School Effects (Standard Deviations) - Broader Sample

	Elementary school		Middle school		High school	
	Healthy students (1)	Unhealthy students (2)	Healthy students (3)	Unhealthy students (4)	Healthy students (5)	Unhealthy students (6)
	Outcome: ELA test scores (broader sample)					
Non-shrunk ($\hat{\theta}_{j\hat{h}}$)	0.114	0.145	0.160	0.159	0.142	0.162
	(0.003)	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)
EB shrunk ($\hat{\theta}^*_{j\hat{h}}$)	0.074	0.094	0.117	0.117	0.094	0.111
	(0.002)	(0.002)	(0.004)	(0.004)	(0.003)	(0.003)
Number of schools	783	783	525	525	570	570
	Outcome: ELA test scores (main analysis sample for reference)					
Non-shrunk ($\hat{\theta}_{j\hat{h}}$)	0.115	0.151	0.148	0.157	0.136	0.163
	(0.003)	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)
EB shrunk ($\hat{\theta}^*_{j\hat{h}}$)	0.066	0.090	0.114	0.109	0.085	0.107
	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)	(0.003)
Number of schools	764	764	496	496	525	525
	Outcome: Math test scores (broader sample)					
Non-shrunk ($\hat{\theta}_{j\hat{h}}$)	0.157	0.176	0.183	0.180	0.127	0.153
	(0.004)	(0.004)	(0.005)	(0.006)	(0.004)	(0.005)
EB shrunk ($\hat{\theta}^*_{j\hat{h}}$)	0.116	0.125	0.137	0.134	0.087	0.102
	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
Number of schools	783	783	526	526	571	571
	Outcome: Math test scores (main analysis sample for reference)					
Non-shrunk ($\hat{\theta}_{j\hat{h}}$)	0.152	0.178	0.176	0.175	0.119	0.147
	(0.004)	(0.005)	(0.006)	(0.006)	(0.004)	(0.005)
EB shrunk ($\hat{\theta}^*_{j\hat{h}}$)	0.103	0.118	0.133	0.123	0.081	0.096
	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)
Number of schools	763	763	497	497	525	525

Notes: This table reports dispersion of non-shrunk and Empirical Bayes shrunk subgroup-specific school effect estimates using a broader sample of students anytime enrolled in Medicaid and anytime eligible for subsidized lunch (as opposed to half-time for each). Columns (1), (3), and (5) are for healthy students; Columns (2), (4), and (6) are for unhealthy students. Standard errors are calculated following Ahn and Fessler (2003), which describe standard errors on estimators of the standard deviation of a normal distribution; this is appropriate in our context because we make normality assumptions in the Empirical Bayes shrinkage procedure, though one could also bootstrap the standard errors. Near-identical results (except school number) indicate robustness of the sample.

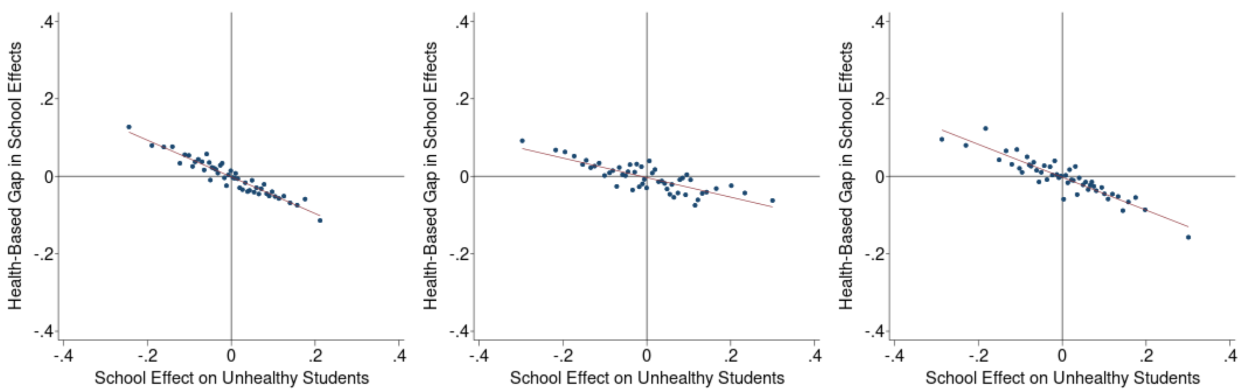
Figure 1E.2: Explaining the Gap in School Effectiveness: Healthy - Broader Sample



(a) Elementary (b) Middle (c) High

Notes: On the x-axis is the estimated school effect on healthy students using ELA test scores as the dependent variable, and on the y-axis is the value-added gap. We use the broader sample of students anytime enrolled in Medicaid and anytime eligible for subsidized lunch (as opposed to half-time for each); similar results indicate robustness of the sample.

Figure 1E.3: Explaining the Gap in School Effectiveness: Unhealthy - Broader Sample



(a) Elementary (b) Middle (c) High

Notes: On the x-axis is the estimated school effect on unhealthy students using ELA test scores as the dependent variable, and on the y-axis is the value-added gap. We use the broader sample of students anytime enrolled in Medicaid and anytime eligible for subsidized lunch (as opposed to half-time for each); similar results indicate robustness of the sample.

1.F Robustness of the Effect of Homebound Teachers

Table 1F.1: Effect of Homebound Teachers on Test Scores
Without the District-Level Spending Per Pupil Control (Equation (1.18))

	Effect on Math test scores		Effect on ELA test scores	
	Healthy students (1)	Unhealthy students (2)	Healthy students (3)	Unhealthy students (4)
	Elementary school students			
Homebound teacher effect	−0.015	−0.003	0.008	−0.003
	(0.021)	(0.031)	(0.020)	(0.029)
Mean test score	−0.352	−0.571	−0.361	−0.562
R-squared	0.553	0.532	0.596	0.574
Student-year observations	95,494	43,780	95,537	43,816
	Middle school students			
Homebound teacher effect	0.110	−0.109	0.251	0.177
	(0.059)	(0.072)	(0.055)	(0.067)
Mean test score	−0.395	−0.728	−0.376	−0.670
R-squared	0.535	0.510	0.586	0.562
Student-year observations	96,839	44,959	96,908	45,047
	High school students			
Homebound teacher effect	0.003	0.060	0.021	0.051
	(0.014)	(0.019)	(0.015)	(0.020)
Mean test score	−0.454	−0.815	−0.433	−0.802
R-squared	0.640	0.557	0.627	0.577
Student-year observations	103,159	43,141	102,942	42,918
School fixed effects	Y	Y	Y	Y
Academic year fixed effects	Y	Y	Y	Y
District-level spending per pupil	N	N	N	N

Notes: This table presents results from the two-way fixed effect regression specified in Equation (1.18) with test scores on the left-hand side but no district-level spending on the right-hand side. All regressions include school and academic year fixed effects, along with controls for student heterogeneity in demographics, health, and ability (past test scores), total school enrollment, and percent unhealthy. District-level spending per pupil is omitted from this specification to show that the main results are not driven by a substitution in resources across students from different health groups (controlling for spending means that in order to hire a homebound teacher, a school or district must take resources from elsewhere). Mean test scores by school type and health group are provided to understand the effect of the homebound teachers in the context of pre-existing differences in test scores.

2 HIGH SCHOOL QUALITY, CURRICULA, AND ACADEMIC ACHIEVEMENT GAPS

2.1 Introduction

Little is known of the characteristics that drive school quality, despite a large body of literature that has explored school quality and its effects on long-term outcomes in adulthood (e.g., Chetty et al., 2014b). In this paper, I first study the relationship between high school quality and the curriculum, focusing in particular on Advanced Placement (AP) course offerings[1] to understand whether one of the reasons some schools are better than others is that they provide more opportunities to their students. I then explore how the availability of advanced courses heterogeneously impacts take-up by students of different demographic characteristics - such as household income, race and ethnicity, English proficiency, and gender - and how advanced course participation has heterogeneous effects on the achievement of students in different demographic subgroups.

Test score disparities among students that are from different socioeconomic subgroups but otherwise have access to the same school inputs are a longstanding issue in the US; moreover, these achievement gaps continue to persist and are widening in many states. The National Center for Education Statistics (NCES) points to shrinking achievement gaps in North Carolina, which makes this state a particularly interesting setting for a study on achievement gaps. Is the state doing anything in particular to shrink achievement gaps that other states can learn from? Or are North Carolina's shrinking achievement gaps being driven by decreasing test scores among the high-achieving group? Understanding the driving factor behind shrinking achievement gaps in North Carolina is an important motivation for this project, and can help both North Carolina as well as other states in setting policy.

To answer my research questions, I use detailed student-level data from the North Carolina Education Research Data Center (NCERDC), which contains demographic characteristics, standardized test scores, and detailed transcript information for over 700 thousand unique students at public high schools across the state from 2012-2018. I use a value-added model to estimate high school quality, following from some of the recent literature (Abdulkadiroğlu et al., 2020). While much of the past literature estimates school quality and takes its inputs for granted, I use a two-way fixed effects design to decompose my measure of quality into three primary components: (1) factors from the curriculum, (2) teacher characteristics, and (3) composition of the student body. I find that the composition of the student body - in particular, incoming achievement from middle school test scores - is the largest driver of quality, indicating that high school quality is dependent on incoming student quality. A higher-achieving incoming peer group might attract better teachers and better resources, which in turn leads to higher school quality.

Moving beyond school quality, I also estimate the impact of advanced course offerings and advanced course *participation* on student achievement, paying special attention to heterogeneous

[1]Though AP courses are a subset of advanced courses in high schools, I refer to the two interchangeably, because other types like International Baccalaureate courses are beyond the scope of this paper.

effects by demographic subgroup. I find that high schools' advanced course offerings have little direct impact on eleventh grade ACT scores, regardless of household income, race and ethnicity, English proficiency, and gender; this corroborates findings from Owen (2024), which finds that the availability of AP courses has little impact on college outcomes, except among the highest-achieving students. As a direct corollary of null effects on student achievement, advanced course offerings also have no impact on achievement gaps. Despite these null results, I find that the availability of advanced courses induces different levels of take-up among students from different groups. Not only do students of higher prior ability (measured from middle school test scores) select into advanced courses, students from typically higher-resource households (White, Asian, and non-economically disadvantaged students more generally) also do. In addition, students from higher-resource households experience the greatest positive effect of participation in advanced courses on their eleventh grade ACT scores, indicating that the students that select into advanced courses are also generally the ones that benefit the most from them.

This paper is closely related to several strands of economic literature. The first strand estimates school or teacher quality (also referred to as effectiveness) through value-added methods (e.g., Wright et al., 1997; Rothstein, 2010; Chetty et al., 2014a; Angrist et al., 2017; Abdulkadiroğlu et al., 2020). Making use of the panel structure of my data that extends for several years, I extend classic models of school value-added to allow for time-variation in school quality. Frequently, researchers use value-added methods to estimate school quality and then identify how exposure to higher quality education in childhood impacts long-term outcomes in adulthood. Few papers have investigated the school inputs that improve quality, which is my main contribution to the value-added literature. By decomposing school quality into factors from the curriculum, the teachers, and the student body, I aim to understand exactly why higher quality schools are actually better.

A second strand of literature explores inequities in access to AP courses. Hockman (1970) describes early expansion of the Advanced Placement system as being concentrated among White students from higher-income households. Schneider (2009) finds that even when Advanced Placement expanded to schools serving students from lower-income households, it was slow and often difficult due to resource constraints. Most similar to my paper is Owen (2024), which investigates how the availability of AP courses in Michigan high schools impacts course take-up and college outcomes for students from different demographic subgroups. Many of the results from my paper corroborate Owen's findings: in particular, we both find that students from higher-income households as well as White and Asian students are more likely to take advantage of advanced courses when offered.[2] I further show how participation has positive impacts on achievement, especially for White students and those from higher-income households. I contribute to this strand of literature by extending previously used methods to the North Carolina setting - which is particularly interesting due to shrinking achievement gaps - and substantiating prior results.

A third strand of the literature looks at the effects of high school curricula on college outcomes.

[2]In this regard, my paper ties into the vast literature on selection (Willis and Rosen, 1979; Heckman et al., 1998; Heckman and Li, 2004), which could be explored further in future work.

Altonji (1995) exploits variation in curricula across US high schools to show that additional years of core courses in high school increase post-secondary education. Aughinbaugh (2012) uses variation in curricula to show that a more rigorous high school math curriculum is associated with a higher probability of college attendance while Jia (2021) finds that stricter high school math requirements increase the proportion of students that go on to get a STEM degree[3] in college. Arce-Trigatti (2018) studies an Arkansas mandate on advanced course offerings, finding that increases in advanced course offerings cause a decrease in overall college enrollment but an increase in the four-year college graduation rate, implying a change in the sorting of high school students into colleges. My paper is also related to Jackson (2010), which finds that a Texas monetary-based AP course incentive program for students and teachers increases advanced course participation and scores, ACT scores, and college matriculation. Much of this body of work concentrates on college outcomes, which restricts the sample to high school students that enter college; this may create a selection issue. I focus on academic outcomes during high school, which allows the sample to include a broader group of students, and will help in understanding whether previously-estimated long-term effects of advanced courses are founded in the short-term.

The remainder of this paper will proceed as follows. Section 2.2 will first outline the data and present basic descriptive statistics, then provide background on the curriculum and North Carolina policy. Section 2.3 lays out a conceptual framework to motivate the high school curriculum as a driver of school quality and describes the mechanisms by which advanced course offerings might generate heterogeneous gains to test scores. For my main empirical analysis, Section 2.4 will outline the model of school value-added and decompose school quality into factors from the curriculum, teacher characteristics, and the student body composition. Sections 2.5 and 2.6 will present two-way fixed effects models to understand the effects of advanced course offerings and participation on student achievement. Finally, Section 2.7 will conclude.

2.2 Data

Data Description and Descriptive Statistics

For this project, I use several individual-level datasets from the North Carolina Education Research Data Center (NCERDC), which are linked together by anonymized personal identifiers. In terms of students' academic outcomes, the data include ACT scores (and grade 3-8 standardized test scores), GPA, and attendance. I also observe a rich set of demographics that includes race/ethnicity, sex, subsidized lunch status, English language proficiency, and disability status. Lastly, I observe student-level transcript information, which includes all course titles, difficulties, and grades received at the end of the semester or year. These are instrumental in constructing a time-varying array of course offerings for each high school. Though many of these data elements go back to the early 2000s or earlier, my primary outcome of interest - ACT scores - are available starting in 2012; for this

[3]STEM degrees include Science, Technology, Engineering, and Mathematics.

reason, much of my analysis is limited to 2012-2018, though I control for grade 8 test scores going back to 2009 for students in grade 11 in 2012.

On the teacher side of the data, I observe educational attainment, experience, pay, and basic demographic information such as race/ethnicity and sex. Importantly, I can link teachers both to the schools at which they teach and the students to whom they teach, the latter of which comes from student-level course membership data. In addition, I have a principal indicator and principal experience, which allow for the use of a principal fixed effect in my analyses; this is important because of the control principals have over decisions regarding the school curriculum. The limitation with the teacher-level data is that I have only full information on teachers in traditional public schools; I am unable to match charter schools to all teachers, so many of the analyses I conduct in this paper are limited to the traditional public schools. Aside from the students and teachers, I also observe district-level financial information on funding broken down by state, federal, and local government contributions as well as DPI-constructed school and district report cards.

Table 2.1: Student and Teacher Descriptive Statistics, 2012-2018

	Traditional Public Schools (1)	Charter Schools (2)
Grade 11 students		
White	0.538	0.676
Female	0.504	0.512
Eligibile for subsidized lunch	0.400	0.150
Limited English proficient	0.030	0.007
Average grade 8 Math score	0.211	0.370
Average grade 8 Reading score	0.214	0.467
Average ACT score	18.4	20.7
Teachers		
White	0.755	0.754
Female	0.774	0.759
Master's degree or higher	0.374	0.150
Average years of experience	13.3	12.6
Number of unique students	717,648	22,089
Number of unique teachers	152,507	3,090
Number of schools	540	74

Notes: Charter schools are included in Column (2) for completeness but will not be included in subsequent analyses. Grade 8 test scores are z-scored by year prior to de-duplication, which is discussed in Appendix 2.A; positive mean scores are due to de-duplication as well as attrition (some students have no ACT score). The scale of each ACT component (and overall score) is 1-36 and the standard deviation hovers around 5 points each year.

Table 2.1 presents descriptive statistics for students in grade 11 as well as teachers for my main analysis window 2012-2018; I include charter schools to give a sense of the distribution across school types, though the subsequent analyses will not include charters due to data limitations. Corroborating the findings of Gilraine et al. (2023) and other prior literature, charter schools in North Carolina have more White students than traditional public schools, fewer economically disadvantaged students (those eligible for subsidized lunch), and fewer students that have limited English proficiency. Surprisingly, the charter schools have teachers that are on average slightly less educated and less experienced. Because many of the charter schools in North Carolina are younger than traditional public schools, these findings imply that new charter schools tend to hire teachers that are newer to the profession.

Background: Standardized Exams

Students in grades 3-8 are given standardized exams every year; starting in the 2012-13 academic year[4] the PreACT is administered each year to all students in grade 10[5] and the ACT[6] is administered each year to all students in grade 11 per North Carolina General Assembly mandate. Since the ACT is a curriculum- and standards-based exam, it is intended to be calendar year-invariant; in other words, it tests whether the standards of the current curriculum are appropriate. As a consequence, the ACT Corporation argues that its exam is comparable across years, a feature that will be important in my estimation of time-varying school value-added.

Background: Curriculum and Advanced Placement Courses

Though there are statewide minimum standards in place for each high school course and minimum requirements for high school graduation, local schools and districts may require additional courses or credits for graduation. This gives high schools some freedom in choosing courses to offer to students (e.g., Advanced Placement courses), which in turn creates variation in curricula across high schools and across time.

The Advanced Placement program started in 1955 in the US to increase college entrants and graduates (Rothschild, 1999). It is run by the College Board, a nonprofit that helps high school students find success in college through services like the SAT and programs like Advanced Placement. The AP program gives high schools the opportunity to offer students a suite of intro-level college classes that meet a national standard set by the College Board. For students, the primary advantage of taking an AP course is the opportunity for college credit contingent on passing the AP course exam, without being weighed down by the price tag of a regular college course. Over the years, participation in the AP program has grown over all demographic groups; Chatterji et al. (2021)

[4]For simplicity, I will refer to an academic year by the fall semester in which it starts. For example, 2012 represents the 2012-13 academic year.

[5]The PreACT provides an early ACT experience so that students know what to expect in grade 11.

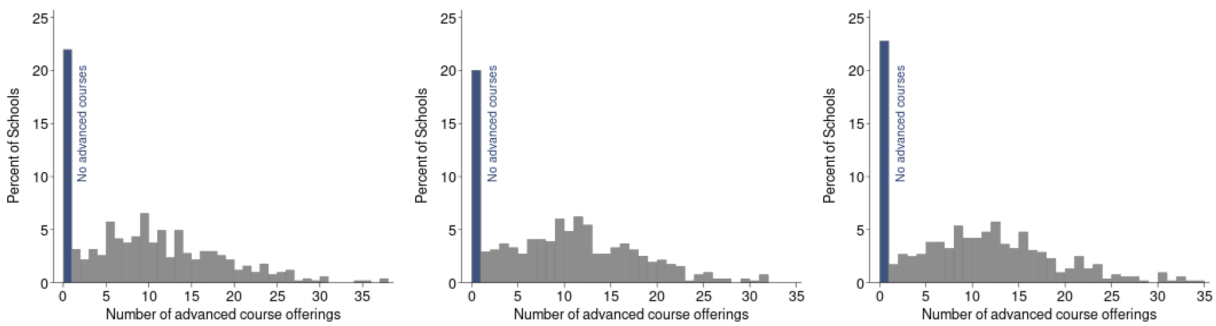
[6]The ACT is a competitor of the SAT; it is used in college admissions and for merit-based scholarships.

find that around 85% of all public high school students nationwide had access to at least one AP course in their high school in 2015.[7]

Variation in the Curriculum

In this section I present some of the data patterns that I find regarding advanced course offerings in North Carolina high schools. Figure 2.1 shows the distribution of the number of advanced course offerings across schools for a selection of different years in my observation window (while the overall shape of the distribution through time is stable, there is a slight upward shift in density over time). On the extensive margin, around 20% of public schools offer no advanced courses in any given year; on the intensive margin, there is a fairly large spread centered around a mean of 10 advanced course offerings.

Figure 2.1: Distributions of Advanced Course Offerings Across Schools, by Year



(a) 2012 (b) 2014 (c) 2016

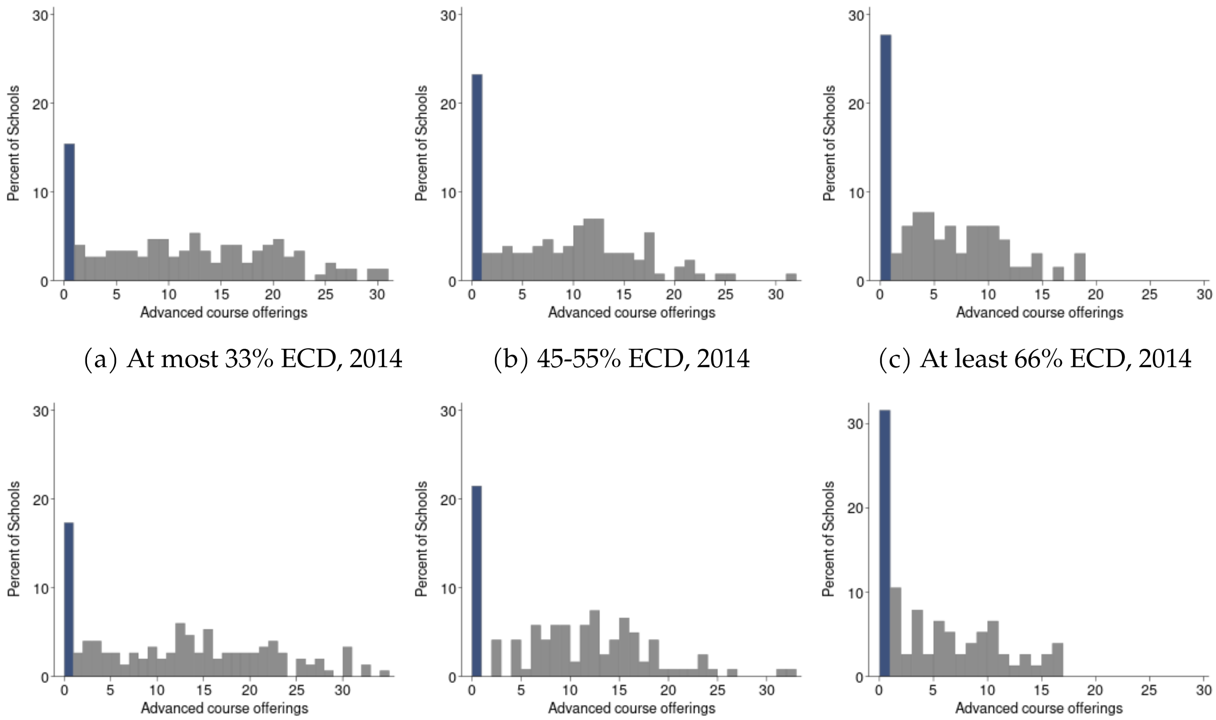
Notes: Each panel shows a histogram of the number of advanced course offerings across schools for a single year (2012 on the left, 2014 in the middle, and 2016 on the right); only traditional public schools are included. Around 20% of public schools offer no advanced courses each year.

Figure 2.2 shows that a substantial portion of variation in advanced course offerings across schools remains even after conditioning on the student body composition - specifically, the percentage of students that are economically disadvantaged.[8] As the proportion of disadvantaged students rises, the likelihood that a school offers no advanced courses increases, the spread of advanced course offerings decreases, and the mean shifts down. More specifically, around 30% of the schools with at least 66% economically disadvantaged students offer zero advanced courses while only 15-20% of schools with at most 33% economically disadvantaged students offer zero advanced courses. However, there is a consistently large distributional spread during my analysis window, which suggests that there is variation in advanced course offerings across schools even when conditioning on the student body composition.

[7]For more background on Advanced Placement, see Owen (2024).

[8]I use the terms "economically disadvantaged" and "eligible for subsidized lunch" interchangeably; both refer generally to students from lower-resource households.

Figure 2.2: Distributions of Advanced Course Offerings,
Conditional on the Student Body Composition



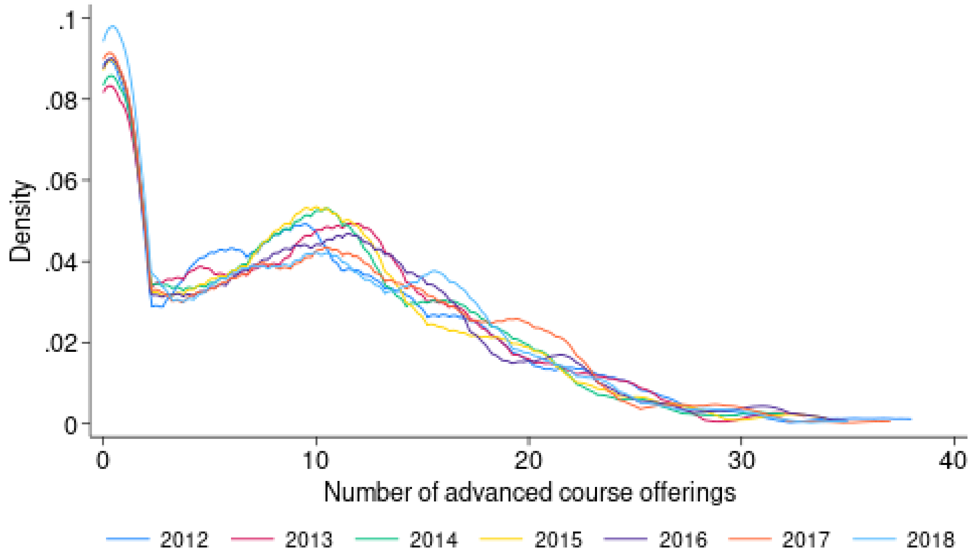
(a) At most 33% ECD, 2014 (b) 45-55% ECD, 2014 (c) At least 66% ECD, 2014

(d) At most 33% ECD, 2016 (e) 45-55% ECD, 2016 (f) At least 66% ECD, 2016

Notes: Each panel shows a histogram of the number of advanced course offerings across schools, conditional on income-based demographic composition of the student body; the left-most panels include schools serving the lowest proportion of economically disadvantaged (ECD) students while the right-most panels include schools serving predominantly disadvantaged students. The top row are for 2014 while the bottom row are for 2016. Only traditional public schools are included.

The previous two figures have focused on variation in advanced course offerings across schools for specific years, but what about over time? While the curriculum is a multifaceted object that likely changes gradually and sometimes requires changes in statewide requirements, the portion of the curriculum that encompasses advanced courses likely varies more from year to year. Figure 2.3 shows the yearly distribution of advanced course offerings across schools from 2012-2018. On the extensive margin, 2012 contains the highest proportion of schools offering no advanced courses; this proportion drops until around 2016 then rises again. There is also variation on the intensive margin, though more work is needed to understand the specific types of schools that have swifter and more frequent changes in their advanced course offerings.

Figure 2.3: Distributions of Advanced Course Offerings Across Time



Notes: This graph depicts variation in advanced course offerings across schools from year to year. On the extensive margin, fewer and fewer schools offer zero advanced courses until around 2016; subsequently the frequency increases. The share of schools in the 5 to 10 range decreases through time; the share of schools above 20 increases through time.

To conclude this section, I turn briefly from the curriculum as an object chosen by the school to participation in the advanced portion of the curriculum - a decision made by the student. Table 2.2 shows the overall demographic composition of two groups over the analysis window (2012-2018): (1) those that participate in advanced courses either in 10th or 11th grade, and (2) those that never take an advanced course. This table provides evidence of selection into advanced courses: students that participate in at least one advanced course have on average around a one standard deviation advantage in incoming Math and Reading scores over students that do not participate in any advanced courses. Demographically, more White and Asian students take AP courses while fewer Black and Hispanic students do; even more striking is how few students in advanced courses are economically disadvantaged compared to the overall fraction of students that are disadvantaged. This provides clear - albeit descriptive - evidence that students from higher-resource households as well as higher-achieving students select into the advanced courses in high school, a point that will be investigated further in the next sections.

Table 2.2: Composition of Student Groups Participating in the Advanced Portion of the Curriculum, 2012-2018

	AP participation in 10^{th} or 11^{th} grade (1)	No AP participation (2)
Group characteristic:		
White	0.672	0.482
Asian	0.059	0.018
Black	0.136	0.299
Hispanic	0.090	0.147
Female	0.576	0.473
Eligible for subsidized lunch	0.189	0.489
Limited English proficient	0.005	0.041
Disabled	0.008	0.106
Average grade 8 Math score	0.928	−0.100
Average grade 8 Reading score	0.888	−0.079
Number of unique students	202,724	513,065

Notes: The first column includes students that took an AP course (i.e., participated in the advanced portion of the curriculum) in either 10^{th} or 11^{th} grade (or in both grades); the second column includes students that took no advanced courses in either grade. I only include here school-year groups for which I observe at least 10 students, which is one of the restrictions imposed in value-added estimation to limit sampling error.

Background: The North Carolina Advanced Placement Partnership

One policy that creates variation in high school curricula is the 2014 North Carolina Advanced Placement Partnership (NCAPP) between the State Board of Education and the College Board. Normally, students pay $97[9] per AP exam; however, the NCAPP requires that the NC Department of Public Instruction (DPI) cover all AP exam fees for public school students that took the associated AP course. On the school side, the NCAPP requires that the NC DPI give prioritized assistance to low-performing school districts that require more attention, support, or resources; in particular, schools are given assistance in rolling out new advanced courses. Annual school report cards are in part dependent upon AP course participation and performance, so a secondary implication of the NCAPP is that schools have incentive to ramp up their AP course offerings.[10]

[9] As of 2022, though this fee increases from year to year.

[10] The NCAPP will be discussed more in Appendix 2.C, since it provides a potential solution to the concern of endogeneity of the curriculum.

2.3 Conceptual Framework

In this section, I will conceptually motivate the high school curriculum as a driver of school quality and student achievement; the following section will then outline the econometric model used to identify the effects of interest. It is well-known in the value-added literature that skill formation in childhood is an important part of K-12 education. As a conceptual motivation for the model to follow, we can think of the current stock of skills - or human capital - that student i who attends school j possesses at time t as a function of previously acquired skills $hc_{i,t-1}$, home inputs H_{it}, and quality of education α_{jt}:

$$hc_{i,t+1} = f_i(hc_{it}, H_{i,t+1}, \alpha_{j,t+1}) \text{ for } t = 0, 1, ..., T-1 \tag{2.1}$$

A child's current period skills, home inputs, and quality of education are indexed by the same time subscript is that I will measure skills by an annual end-of-year standardized exam; in this regard, children are exposed to certain resources at home and in school for nearly an entire school year before their skills for that year are measured. Although Equation (2.1) outlines development of childhood skills from time $t = 0$ (birth) to time $t = T$ (the typical high school graduation age of 18), I focus on skill formation during high school due to its relevance to my research question. Additionally, I have written the f function with an i subscript to emphasize the potentially heterogeneous effects of home and school inputs on skill formation. Lastly, though the home input is an important part of development, I focus on the school side for this analysis.

Breaking into the black box of school quality, I will think about a decomposition of quality into the three components: (1) factors from the curriculum, (2) aggregate teacher input, and (3) composition of the student body, as follows:

$$\alpha_{jt} = g(C_{jt}, T_{jt}, B_{jt}) \tag{2.2}$$

Thinking of a school's curriculum as the array of courses it offers, it is not surprising that all demographic subgroups benefit homogeneously from the same curricular structure. Consider the following two-student example: student i_1 has low eighth grade scores and expects not to participate in advanced courses; student i_2 has high eighth grade scores and expects to take advanced courses. By construction, the second student will benefit more than the first student from a curriculum that offers more advanced courses because these are courses this student will take; furthermore, the first student may see a detriment due to school resources lost to investment in advanced courses. Explicitly, we can express this heterogeneity in the effect of the curriculum as follows:

$$\frac{\partial^2 f_{i_1}}{\partial \alpha_{jt}\, \partial C_{jt}} \neq \frac{\partial^2 f_{i_2}}{\partial \alpha_{jt}\, \partial C_{jt}} \tag{2.3}$$

Now further suppose student i_2 is in demographic subgroup s_2 and add in student i_3 with high eighth grade scores from demographic subgroup s_3. Both students i_2 and i_3 have high incoming test scores and are likely to select into the more advanced high school courses, as suggested by some of the previous literature on education and selection (Arce-Trigatti, 2018; Heckman and Li, 2004). However, not only would students i_2 and i_3 receive different benefits from the same advanced course (as in any course, there would be a distribution of material learned across students), they may actually be incapable of equal gains ex-ante due to differences in household budget constraints, resources, etc. This example then shows that the curriculum may not just heterogeneously affect each of the three students based on who selects into which course, there may also be pre-existing differences between students that limit their potential gains from advanced courses.

To identify the demographic-based heterogeneity in the effect of the curriculum, I separate students into subgroups based on observables: S contains students unlikely to participate in advanced courses (typically these are students from lower-resource households) and S' contains students likely to take advanced courses (typically these are students from higher-resource households). In principle, these groups can be divided by any demographic categorization like income, race/ethnicity, English proficiency, etc. Reduced-form analyses that condition on student type will help in understanding the relationship between the high school curriculum and achievement as well as the demographic-based heterogeneity in effects of advanced course offerings. I will take a similar approach in looking at demographic-based heterogeneity in gains from advanced course participation as well as selection into advanced courses. These econometric models will be discussed further in the following sections.

2.4 Estimating School Quality and Identifying its Inputs

Model of School Value-Added

Following from Abdulkadiroğlu et al. (2020), I take a simple model of school value-added (VA) and add a time component in order to identify the time-varying quality of education at high school j. The formal econometric model is written as follows:

$$Y_{it} = \alpha_{jt} + X_{it}'\beta_X + \varepsilon_{it} \tag{2.4}$$

Indexed by student i and time t in academic years, I use overall ACT test scores as the left-hand side variable Y_{it}.[11] Because the ACT is designed to be curriculum- and standards-based, I argue that this is a time-invariant outcome; for this reason, I do not use z-scores within years on the dependent variable to allow for a more comparable measure of quality across years. Following from Abdulkadiroğlu et al. (2020), I include in the covariate vector X_{it} student race/ethnicity, sex, subsidized lunch status, and eighth grade Math and Reading test scores. Because the eighth grade exams vary in scoring across years, I do use z-scores here.[12] The α_{jt} term represents the time-varying average effect of high school j on students' eleventh grade ACT scores. The error term ε_{ijt} reflects variation in ACT scores that is not captured by X_{it} or the school-year fixed effect.

[11] Robustness using ACT Math scores and ACT Reading scores are presented in Appendix 2.B.

[12] To calculate z-scores, I use each year's statewide mean and standard deviation to normalize the yearly distribution of scores so that they have mean zero and standard deviation one.

I de-mean the school-year fixed effect estimates relative to the 2012 mean, which yields the effect E_{jt} of school j in year t relative to the mean school in 2012. Prior to de-meaning, estimates for year t are centered around the mean ACT score for that year. Re-centering all estimates about the same mean is valid due to the time-invariant nature of the ACT: because the ACT is intended to test the same curricular standards and the same skills across years, any differences in school value-added across years can be attributed to changes in school quality. Moreover, this allows for a simple comparison of school quality from year to year.

$$E_{jt} = \alpha_{jt} - \frac{1}{J}\sum_{j}\left[\alpha_{j,2012}\right] \tag{2.5}$$

The ACT is a college entrance exam that aims to test college and career readiness through a combination of math, science, reading comprehension, writing, and English/vocabulary; it has been used as one of the primary metrics by which US colleges and universities make admission decisions. And because the main focus of every US high school is to prepare its students either for college or career, the holistic approach of the ACT makes it a good proxy for student learning in high school. Therefore, the value that a high school adds to ACT scores represents a measure of how successful that school is in contributing to student learning. For this reason, I argue that the E_{jt} estimated above through a value-added framework are a plausible measure of high school quality.

Underlying Assumption: Selection on Observables

The critical assumption underlying this model is selection on observables: the covariate vector X_{it} is sufficiently rich for potential ACT scores Y_{ijt} to be independent of the school assignment of student i at time t. In other words, the determinants of current test scores are entirely observed, given the school assignment. Formally,

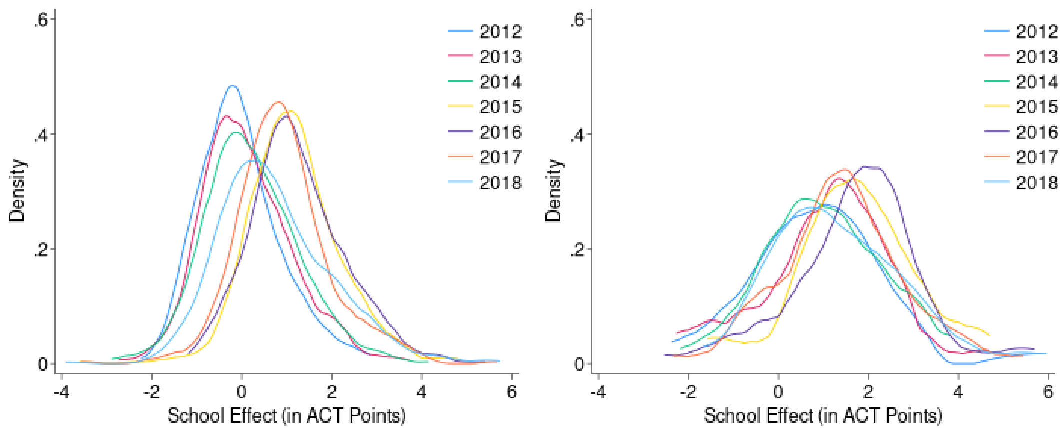
$$E[\varepsilon_{it} \mid j, t, X_{it}] = 0, \quad j = 1, ..., J \tag{2.6}$$

Under this assumption, a standard ordinary least squares (OLS) regression of the fully-interacted model in Equation (2.4) yields unbiased estimates for α_{jt} and β_X. Though there has been criticism of the selection-on-observables assumption (Rothstein, 2010, 2017) and an ongoing debate regarding validity of value-added models, recent literature has found bias in value-added estimates to be limited when compared to rank-ordered control function methods (Abdulkadiroğlu et al., 2020), and in models that control for students' past test scores (Chetty et al., 2014a). Furthermore, (Angrist et al., 2016) find that school policies based on value-added estimates are likely to boost student achievement even if the underlying estimates are not perfectly unbiased.

School Quality Estimates

Figure 2.4 shows yearly distributions of school quality estimates for traditional public schools and charter schools. Charter schools are included to highlight an interesting pattern: the distribution of school effect estimates is consistently much wider among charter schools than traditional public schools, a finding that is robust to using ACT Math or Reading scores as the dependent variable in Equation (1.4) instead over overall ACT scores. This may in part be due to sample size: with only 74 charter schools compared to 540 traditional public schools, the charter school distributions are more prone to outliers. However, Gilraine et al. (2023) point to more variation in charter schools' curricula creating more dispersion in value-added. Without data on charter schools' teachers, this paper is unable to verify this finding; nevertheless, it provides an interesting motivation for an investigation into the curriculum, even if only for the traditional public schools.

Figure 2.4: Distributions of School Effects, by Year



(a) Traditional public schools only

(b) Charter schools only

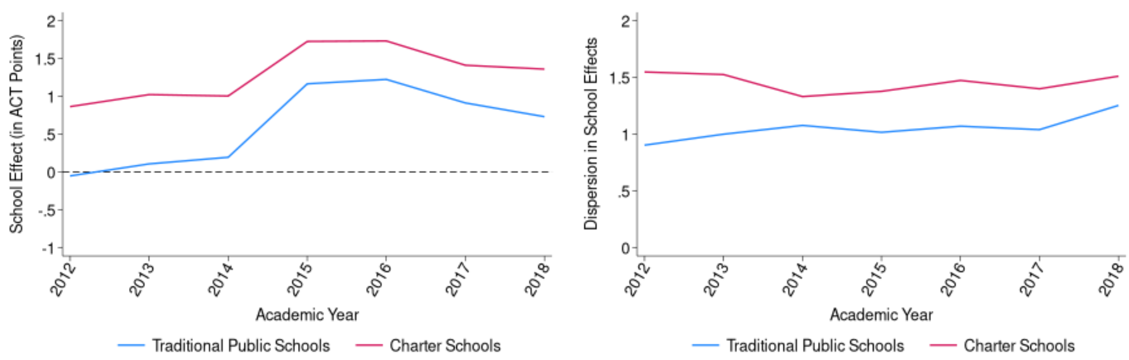
Notes: Each graph shows distributions of de-meaned school effect estimates; the dependent variable is overall ACT score. The left panel includes only traditional public schools while the right panel includes only charter schools. Charter schools are included to show distributional variation across school types and motivate the curriculum as one reason behind this, but they are omitted from subsequent analyses.

Figure 2.5 shows trends in mean school quality (left) and dispersion in school quality (right).[13] From panel (a), it is clear that there is a quality gap between the average charter school and the average traditional public school that is fairly constant over time. The exception is from 2014 to 2015, in which a relatively stable 1 standard deviation (SD) gap shrinks to roughly 0.5 SD; this may be related to a positive impact of the NCAPP (discussed in Section 2.2) on the public school curriculum. The timing would make sense, given the NCAPP's start in 2014, and it likely had a larger effect on tradition public schools because they were tied directly to the NC DPI for funding and standards. From panel (b), dispersion in charter school quality is consistently higher than dispersion in traditional public school quality, which may in part be due to more variation in charter school curricula as a result of more freedom to set standards. While public schools see a steady increase in dispersion over time, there is an initial decrease in dispersion in charter school quality.

[13]Dispersion is defined as the standard deviation of the school effect distribution.

This could be related to the 2012 lifting of North Carolina's statewide charter school cap that Gilraine et al. (2023) study; essentially, there was an explosion of new charters in a short period, which likely caused a higher concentration of near-average-quality charter schools.

Figure 2.5: Trends in School Effect Mean and Dispersion Over Time



(a) Mean (b) Dispersion

Notes: Panel (a) shows trends in value-added of the mean traditional public school (blue) and charter school (red); panel (b) shows trends in the dispersion in value-added across schools. The horizontal dashed line in panel (a) represents the 2012 mean value-added across all schools.

School Quality Decomposition

I now take the school quality estimates E_{jt} and decompose them into three components: (1) factors from the curriculum, (2) teacher characteristics, and (3) composition of the student body. This will help in understanding what contributes to a school's quality (i.e., what makes a better school better), something that has often gone unstudied in past literature. To that end, I project the school effects onto several characteristics that vary across schools and through time:

$$E_{jt} = \beta_0 + C'_{jt}\,\beta_C + \bar{T}'_{jt}\,\beta_T + \bar{B}'_{jt}\,\beta_B + \rho_p + \theta_j + \omega_t + \varepsilon_{pjt} \tag{2.7}$$

I use C_{jt} to be a vector of curricular characteristics; I include number of AP course offerings by subject, percentage of courses that are advanced, and the school average student-teacher ratio. $\bar{T}_{jt}$ includes school-aggregate characteristics of teacher composition, such as education, experience, and demographic make-up. I use $\bar{B}_{jt}$ to represent the student body; I include demographic composition of the school's students and average incoming test score. The time fixed effect ω_t absorbs any factors constant within academic years while the school fixed effect θ_j will absorb any school-level factors constant across time; because past work has pointed to a large variation in principal quality and their importance in managing the teacher workforce (Branch et al., 2012), I include a principal fixed effect ρ_p. This decomposition of a school's quality into factors from the curriculum, teacher effect, and student body composition sheds light on what drives school quality, and helps in understanding the role of the curriculum. An interacted model would allow for an investigation into the complementarities between teachers and the curriculum, though I find little here.

Table 2.3: Decomposition of High School Quality (Equation (2.7))

	Unweighted		Weighted
	(1)	(2)	(3)
Curriculum:			
Number AP Math courses	−0.045	−0.023	−0.016
	(0.032)	(0.027)	(0.026)
Number AP Science courses	0.012	0.014	0.021
	(0.021)	(0.016)	(0.015)
Number AP Art courses	−0.034	−0.004	0.001
	(0.028)	(0.021)	(0.021)
Number AP History courses	−0.059	−0.015	−0.005
	(0.022)	(0.019)	(0.018)
Number AP English courses	−0.159	−0.002	−0.019
	(0.042)	(0.039)	(0.036)
Number AP Language courses	0.104	−0.029	−0.025
	(0.032)	(0.022)	(0.020)
Student-teacher ratio	0.015	−0.009	−0.002
	(0.006)	(0.018)	(0.016)
Aggregate teacher characteristics:			
Percent with a master's degree or higher	0.099	−0.011	−0.007
	(0.011)	(0.011)	(0.009)
Average years of experience	0.015	0.018	0.015
	(0.011)	(0.015)	(0.013)
Student body composition:			
Average grade 8 Math score	−0.974	−0.814	−1.043
	(0.196)	(0.351)	(0.288)
Average grade 8 Reading score	1.553	1.344	0.617
	(0.222)	(0.407)	(0.372)
Percent qualifying for subsidized lunch	−0.012	0.006	0.004
	(0.002)	(0.002)	(0.002)
Percent male	0.001	0.009	0.006
	(0.005)	(0.009)	(0.008)
Percent white	−0.012	0.009	0.012
	(0.002)	(0.008)	(0.008)
Principal fixed effects		Y	Y
School fixed effects		Y	Y
Academic year fixed effects		Y	Y
R-squared	0.422	0.872	0.897
School-year observations	3332	3112	3112

Notes: The dependent variable is value-added estimated with overall ACT scores. Weighted specifications weight by total school enrollment; standard errors are clustered at the school level. Only traditional public schools are included in the analysis.

Table 2.3 presents results from the quality decomposition. From Column (1), observable time-varying factors in the curriculum, teacher workforce, and student body explain over 40% of the variation in school quality; nearly 50% of the variation is explained by unobserved factors picked up by the principal, school, and academic year fixed effects, as indicated by the jump in R-squared from Column (1) to Column (2). The finding that unobservable factors explain much of the variation in school quality is unsurprising: part of the reason for a shortage of literature exploring what goes into school quality is a result of data limitations (only so many school characteristics are observable) and econometric difficulty (little variation in school quality, especially over time).

From the results in Table 2.3, most of the observable impact on school quality can be attributed to the student body composition, specifically incoming grade 8 test scores. This suggests that a higher-achieving incoming peer group attracts better teachers and more resources, leading to higher school quality. Recall that one of the key findings of Abdulkadiroğlu et al. (2020) is that when parents are able to choose the school for their kid, they do not value school effectiveness, but rather care about achievement of the potential peer group. My finding indicates that school quality is itself driven in large part by the achievement of the incoming students, specifically achievement tied to reading comprehension and communication skills. This would suggest that - if parents had a say in children's school assignment and valued in a school what they are able to observe (incoming peer group) - they would end up making a decision based on factors highly predictive of school quality.

2.5 Advanced Course Offerings and Student Achievement

I now turn from an exploration of the curriculum in the scope of school quality to an investigation into how the curriculum impacts student achievement. I am interested in understanding whether different curricula are better suited to different student subgroups. I define two sets of student subgroups based on observable characteristics: S to represent groups from lower-resource households and S' to represent groups that are typically from higher-resource households.

$$S = \{\text{economically disadvantaged, Hispanic (Black), limited English proficient}\}$$

$$S' = \{\text{not economically disadvantaged, White, English proficient}\}$$

The first index in each set S and S' represents subgroups divided by household income, which is determined by subsidized lunch eligibility; the second index represents subgroups divided by race and ethnicity (in S I look at both Hispanic and Black student subgroups in comparison with the White student subgroup); the third index divides students by English proficiency. The within-school pairwise difference between the average outcome of subgroups S'_k and S_k is the achievement gap between those two groups.

One option to understand the effect of changes in the curriculum on achievement gaps is to project within-school achievement gaps between different student subgroups on a vector of curricular variables; however, this would limit variation of the outcome variable to the school level rather than the student level. Instead, I take advantage of the NCERDC's individual-level data by placing students' ACT scores directly in the model:

$$\begin{aligned} Y_{it} = \delta_{0,k} + \delta_{S_k}[\mathbb{1}(i \in S_k) \times C_{jt}] + \delta_{S'_k}[\mathbb{1}(i \in S'_k) \times C_{jt}] \\ + X'_{it}\,\delta_{X,k} + \bar{T}'_{jt}\,\delta_{T,k} + \rho_p + \eta_{js} + \omega_t + \varepsilon_{it} \quad \text{for } k = 1, 2, 3 \end{aligned} \tag{2.8}$$

For the left-hand side variable, I use ACT Math scores and ACT Reading scores to understand how changes in advanced course offerings impact performance on different parts of the college entrance exam. C_{jt} is the number of advanced course offerings available at school j at time t; I include separate specifications that look at Math courses and English (or Reading) courses as the primary curricular variable. X_{it} includes student demographics and past test scores; $\bar{T}_{jt}$ includes the school-average teacher characteristics of education, experience, and demographics.

$k \in \{1, 2, 3\}$ represents the subgroup category index where $i \in S_1$ means student i is economically disadvantaged and $i \in S'_1$ means student i is not economically disadvantaged. The idea is to estimate Equation (2.8) separately for each subgroup index k, which will yield an estimate of δ_{s_k} for each lower-resource subgroup in S and an estimate of $\delta_{s'_k}$ for each higher-resource subgroup in S'_k. These represent subgroup-specific effects of changes in advanced course offerings on ACT scores. I also include principal, school-subgroup, and academic year fixed effects. The school-subgroup fixed effect is important because it removes any between-subgroup variation from the estimate of δ, both within and between schools.

Comparing the δ_{s_k} and $\delta_{s'_k}$ between corresponding subgroups in S and S' will shed light on the effect of changes in advanced course offerings on achievement gaps. Take for example the income-based student subgroup division, and let the C_{jt} be advanced math courses. Then

$$\begin{cases} \hat{\delta}_{S_1} < \hat{\delta}_{S'_1} & \implies \text{adding an advanced Math course widens the income-based gap} \\ \hat{\delta}_{S_1} > \hat{\delta}_{S'_1} & \implies \text{adding an advanced Math course shrinks the income-based gap} \end{cases}$$

Table 2.4 presents results for subgroups determined by subsidized lunch eligibility (top), race and ethnicity (middle), and English proficiency (bottom). Overall, I find null effects of changes in both advanced Math course offerings as well as advanced English course offerings on ACT scores.[14] There are a few reasons for null effects across the board. The first is that ACT scores are not the ideal outcome here. Because the ACT is a standards-based exam that tests the minimum standard required to produce the skills necessary to succeed, changes in advanced courses that focus on specialized or more technical skills might not improve ACT performance because the students taking these courses are likely already performing well above the minimum standard. Another possibility is that there is enough balance of advanced course participation within each demographic subgroup that any effects on the group that selects into advanced courses are neutralized by opposite effects on the group that select out of advanced courses.

[14]Though not statistically significant, the effects are more generally favorable among the lower-resource groups, which suggests minor improvements to achievement gaps.

Table 2.4: Effect of Advanced Course Offerings on Student Subgroups (Equation (2.8))

	Effect on ACT Math scores		Effect on ACT Reading scores	
	Effect of a change in AP course offerings in			
	Math (1)	English (2)	Math (3)	English (4)
Economically disadvantaged	−0.011	−0.046	0.017	0.022
	(0.023)	(0.035)	(0.030)	(0.047)
Group mean outcome score	16.9	16.9	16.3	16.3
Not economically disadvantaged	−0.031	−0.045	−0.037	−0.044
	(0.021)	(0.032)	(0.025)	(0.044)
Group mean outcome score	20.2	20.2	20.4	20.4
R-squared	0.681	0.681	0.568	0.568
Student-year observations	568,574	568,574	568,177	568,177
Hispanic	0.004	−0.037	0.055	0.044
	(0.030)	(0.042)	(0.043)	(0.070)
Group mean outcome score	17.4	17.4	16.7	16.7
Black	−0.036	−0.047	−0.013	0.051
	(0.025)	(0.035)	(0.034)	(0.058)
Group mean outcome score	16.5	16.5	15.7	15.7
White	−0.021	−0.030	−0.024	−0.054
	(0.023)	(0.034)	(0.026)	(0.047)
Group mean outcome score	20.3	20.3	20.6	20.6
R-squared	0.685	0.685	0.571	0.571
Student-year observations	568,158	568,158	567,761	567,761
Limited English proficient	0.046	0.051	0.112	0.065
	(0.047)	(0.075)	(0.085)	(0.142)
Group mean outcome score	15.1	15.1	12.8	12.8
English proficient	−0.025	−0.049	−0.016	−0.022
	(0.020)	(0.031)	(0.023)	(0.041)
Group mean outcome score	19.0	19.0	19.0	19.0
R-squared	0.678	0.678	0.567	0.567
Student-year observations	568,531	568,531	568,134	568,134
Fixed effects	Y	Y	Y	Y

Notes: This table presents results from the regression specified in Equation (2.8) for student subgroups; the top uses income-based groups, the middle uses race/ethnicity-based groups, and the bottom uses English proficient-based groups. Columns (1) and (3) use changes in advanced Math course offerings in the right-hand side curriculum variable C_{jt} while Columns (2) and (4) use changes in advanced English courses; Columns (1) and (2) use ACT Math scores as the left-hand side variable while Columns (3) and (4) use ACT Reading scores. Economically disadvantaged students are those that qualify for subsidized lunch. All regressions include principal, school, and academic year fixed effects, along with controls for student characteristics and teacher characteristics. Standard errors in parenthesis are clustered at the school level. Mean test scores are provided to understand effects in the context of pre-existing differences in test scores.

2.6 Advanced Course Participation and Student Achievement

In this section I study the direct effects of student participation in advanced courses rather than the school's impact on students through advanced course offerings. To this end, I construct two individual-level "treatment" variables: T_i^{10} for whether student i took at least one advanced course in grade 10, and T_i^{11} for whether student i took at least one advanced course in grade 11. I then project overall ACT scores on these treatment variables to find the impact of advanced course participation on college and career readiness, as follows:

$$Y_{it} = \lambda_0 + \lambda_1 T_i^{10} + \lambda_2 T_i^{11} + \lambda_3 [T_i^{10} \times T_i^{11}] + X'_{it}\,\lambda_X + \theta_j + \omega_t + \varepsilon_{it} \tag{2.9}$$

In Equation (2.9), λ_1 estimates the effect of grade 10 participation in advanced courses on following-year ACT scores while λ_2 estimates the effect of grade 11 participation. The interaction term is important because it will tell whether there are complementarities between the timing of advanced course participation. Following the same convention as previous sections, included in the X_{it} vector are student characteristics like race/ethnicity, subsidized lunch status, sex, and past Math and Reading test scores; in addition, I include school fixed effects θ_j and academic year fixed effects ω_t to isolate the within-school variation.

I extend this model to include interactions between the advanced course participation variables and student subgroup indicators, to identify whether students from some demographic subgroups experience a larger benefit from taking advanced courses. Formally:

$$\begin{aligned} Y_{it} = \lambda_{0,k} &+ \underbrace{\lambda_{S_k}^{10}[\mathbb{1}(i \in S_k) \times T_i^{10}] + \lambda_{S'_k}^{10}[\mathbb{1}(i \in S'_k) \times T_i^{10}]}_{\text{subgroup-specific effects of }10^{\text{th}}\text{ grade AP participation}} \\ &+ \underbrace{\lambda_{S_k}^{11}[\mathbb{1}(i \in S_k) \times T_i^{11}] + \lambda_{S'_k}^{11}[\mathbb{1}(i \in S'_k) \times T_i^{11}]}_{\text{subgroup-specific effects of }11^{\text{th}}\text{ grade AP participation}} \\ &+ \lambda_{S_k}^{10,11}[\mathbb{1}(i \in S_k) \times T_i^{10} \times T_i^{11}] + \lambda_{S'_k}^{10,11}[\mathbb{1}(i \in S'_k) \times T_i^{10} \times T_i^{11}] \\ &+ X'_{it}\,\lambda_{X,k} + \eta_{js} + \omega_t + \varepsilon_{it} \quad \text{for } k = 1, 2, 3 \end{aligned} \tag{2.10}$$

The first term in each row has an indicator for the historically lower-resource subgroup (economically disadvantaged, Hispanic, Black, and limited English proficient) while the second term in each row has an indicator for the higher-resource subgroup (not economically disadvantaged, White, and English proficient). Analogous to Equation (2.8), $k \in \{1, 2, 3\}$ represents the subgroup category index and the idea is again to estimate the model separately for each subgroup index k, which will yield separate estimates of advanced course participation for each student subgroup. Comparing estimates of λ_{S_k} with $\lambda_{S'_k}$ for each k will shed light on whether there are heterogeneous effects by demographic subgroup.

Table 2.5: Effects of Advanced Course Participation (Equation (2.9))

Advanced course participation in grade 10	1.439
	(0.143)
Advanced course participation in grade 11	1.610
	(0.034)
Interaction term	0.048
	(0.150)
School fixed effects	Y
Academic year fixed effects	Y
R-squared	0.701
Student-year observations	591,794

Notes: The dependent variable is overall ACT score; school and academic year fixed effects are included. Standard errors in parenthesis are clustered at the school level.

Table 2.5 shows the estimated average effect of participating in at least one advanced course in grade 10 or 11. Despite null effects of advanced course offerings, there is a substantial benefit of participation in advanced courses. Furthermore, there is a slight compounding effect of consistent participation, as indicated by a positive interaction term.

The subgroup analysis in Table 2.6 highlights substantial gains from advanced course participation for nearly every student subgroup. Moreover, the effects are consistently and substantially larger for the higher-resource groups, particularly the group of students that are White or not economically disadvantaged. The complementarity between grade 10 and grade 11 advanced course participation is positive for the students from lower-resource groups, indicating that these students see larger bumps in ACT scores when consistently participating in advanced courses. Overall, while I find null effects of advanced course offerings in schools, I do find that the students choosing to participate in advanced courses receive large gains to ACT scores. This would support the story that positive effects of advanced course offerings on students participating in advanced courses are neutralized by null or even negative effects on students choosing not to participate. For the students that do not participate in advanced courses, we would expect null effects of an advanced course offering unless schools are resource-constrained. If this were the case, schools may need to divert resources away from lower-achieving students in order to offer advanced courses to higher-achieving students, which could actually generate a negative effect of advanced course offerings on students that select out of them.

Table 2.6: Subgroup Effects of Advanced Course Participation (Equation (2.10))

Subgroups determined by:	Income (1)	Race/Ethnicity (2)	(3)	English proficiency (4)
	Effect on relevant lower-resource subgroup			
	Economically disadvantaged	Hispanic	Black	Limited English proficient
Participation in grade 10	0.765	0.921	0.454	−0.271
	(0.079)	(0.113)	(0.101)	(0.412)
Participation in grade 11	1.083	1.194	0.702	0.270
	(0.034)	(0.039)	(0.034)	(0.129)
Interaction term	0.209	0.121	0.496	0.497
	(0.087)	(0.126)	(0.113)	(0.718)
	Effect on relevant higher-resource subgroup			
	Not economically disadvantaged	White		English proficient
Participation in grade 10	1.698	1.872		1.167
	(0.160)	(0.161)		(0.135)
Participation in grade 11	1.829	1.999		1.207
	(0.035)	(0.035)		(0.034)
Interaction term	−0.221	-0.487		0.221
	(0.163)	(0.161)		(0.143)
Fixed effects	Y	Y		Y
R-squared	0.705	0.699		0.702
Student-year observations	591,774	545,129		591,729

Notes: The dependent variable is overall ACT score; school-subgroup and academic year fixed effects are included in all specifications. The top half of the table presents effects on lower-resource student subgroups while the bottom half presents effects on higher-resource student subgroups. Standard errors in parenthesis are clustered at the school level.

Selection

There is a well-documented selection of students into high school courses based on prior ability: students with higher test scores in grade 8 are substantially more likely to take advanced courses in high school, especially due to tracking in some of the course subjects. For this reason - to avoid upward bias generated by higher-ability students selecting into advanced courses - all models in this paper have controlled for past test scores. Less clear or obvious is if students select into advanced courses based on observable characteristics other than past test scores, like race/ethnicity, sex, disability status, or other demographics. I test for this selection as follows:

$$T_i^g = \lambda_0^{g,\,\text{select}} + X_{it}'\,\lambda_1^{g,\,\text{select}} + \theta_j + \omega_t + \varepsilon_{it} \quad \text{for } g \in \{10,\ 11\} \tag{2.11}$$

In separate specifications, I project an indicator for whether or not student i took an advanced course in grade 10 or grade 11 (these are the previously-constructed "treatment" variables T_i^{10} and T_i^{11}) on observable characteristics of high school students. As with previous models, included in the X_{it} vector are individual-level demographic characteristics as well as past test scores; I also include in the model school and academic year fixed effects. The estimates contained in the $\lambda_1^{10,\text{ select}}$ and $\lambda_1^{11,\text{ select}}$ coefficients will shed light on some of the observable dimensions by which students select into advanced courses.

Table 2.7: Selection into Advanced Courses (Equation (2.11))

	(1) Participation in grade 10	(2) Participation in grade 11
Incoming grade 8 Math score	0.085	0.145
	(0.004)	(0.003)
Incoming grade 8 Reading score	0.046	0.096
	(0.002)	(0.002)
Economically disadvantaged	−0.031	−0.074
	(0.002)	(0.002)
Race/ethnicity (relative to the White subgroup)		
Asian	0.131	0.116
	(0.017)	(0.008)
Black	−0.019	−0.027
	(0.004)	(0.004)
Hispanic	−0.020	−0.015
	(0.003)	(0.004)
Disabled	0.030	0.016
	(0.003)	(0.001)
Limited English proficiency	0.038	0.049
	(0.005)	(0.006)
Male	−0.028	−0.071
	(0.002)	(0.002)
Fixed effects	Y	Y
R-squared	0.273	0.372
Student-year observations	629,985	629,985

Notes: The dependent variable is overall ACT score; school and academic year fixed effects are included in both specifications. Standard errors in parenthesis are clustered at the school level. Each row gives the marginal probability of participation in advanced high school courses in each grade.

Table 2.7 shows the resulting estimation of Equation (2.11). As expected, the students with higher incoming grade 8 Math and Reading scores were far more likely to select into advanced courses: a one standard deviation increase in prior Math scores amounts to a 9% increase in the likelihood of advanced course participation in grade 10, and a 15% increase in likelihood of participation in grade 11. Asian students are substantially more likely to take advanced courses than White students while Black and Hispanic students are slightly less likely to take advanced courses. Economically disadvantaged students are also less likely to select into advanced courses. Somewhat surprisingly, students that are limited in English proficiency are actually *more* likely to take advanced courses than those that are proficient; this may be a separate selection issue altogether regarding which limited English proficient students attend public high schools in North Carolina as opposed to being home-schooled. Altogether - if we combine the results of Table 2.6 with the results of Table 2.7 - the students that generally receive the largest benefit of advanced course participation (non-economically disadvantaged students and White students) are also the students that are more likely to select into advanced courses in the first place.

This would imply a Roy-type selection underlying advanced course participation: students that expect to receive the largest gain from advanced courses are the ones that self-select into them. Some high school students - particularly the higher-achieving ones - are forward-looking in the sense that they make decisions based on what will get them into the best college or university; under the assumption that all of the top schools place a heavy weight on advanced courses when determining admissions, it is reasonable to think that these students are selecting into advanced courses based on expected gains. Guided by Heckman et al. (1998) and Willis and Rosen (1979), this selection is something that future research could further characterize.

2.7 Conclusion

This paper first studies how changes in the high school curriculum impact school quality. In particular, I focus on advanced course offerings to understand whether schools generate test score gains by offering more opportunities to the high-achieving students. This paper then investigates how advanced course offerings - a decision made by the school - impact achievement gaps, and how participation in advanced courses - a decision made by the student - impact student achievement. Using individual-level data on students and teachers from the North Carolina Education Research Data Center, I use value-added methods to estimate school quality followed by a two-way fixed effects design to estimate the impacts of advanced course offerings and participation.

I find that the composition of the student body is the largest factor driving school quality, indicating that high school quality is highly dependent on incoming student quality. This could indicate that better incoming peers attract schools to hire better teachers and expend more resources, which in turn boosts school quality. Neither the curriculum nor aggregate teacher characteristics have much influence on high school quality. Turning to student achievement more directly, I find that high schools' advanced course offerings have little direct impact on eleventh grade ACT scores or on achievement gaps. Despite these null results, I find that the availability of advanced courses induces different levels of take-up among students from different groups. Not only do students of higher prior ability select into advanced courses, students from typically higher-resource households (White and non-economically disadvantaged students) also do. In addition, students from higher-resource households experience the largest test score gains as a result of participation in advanced courses, indicating that the students that select into advanced courses are also generally the ones that benefit the most from them.

There is much potential for future research in this area. The curriculum is a particularly understudied object, and there are several interesting facets to investigate, relating to high school and college courses and policies. More work is needed to reconcile the finding of null effects of advanced course offerings with strong positive effects of advanced course participation. Specifically, it would be interesting to understand whether the marginal student selecting out of advanced courses experiences a null effect of an advanced course offering or a negative effect due to a shifting of school resources. In addition, the analysis from this paper can directly be extended to look at longer-term outcomes, with a link between high school data from the North Carolina Education Research Data Center to college data from the National Student Clearinghouse. There is also much to be done in quantifying selection into advanced courses and extending this to investigate how college students select into courses and majors.

2.A Dropping Duplicated Test Score Observations

For the main analysis of this paper, I keep only valid test scores and restrict to one test score observation per student per year. In the NCERDC data, there are some students marked as having multiple test score observations within a single year;[15] for these students, I take an average of their test scores within a single year. This ends up only being a concern for the eighth grade test scores of some students in 2009-2011; as shown in Table 2A.1, there are extremely low numbers of person-year test score duplicates in both the ACT and eighth grade exams for 2012 onward.

Table 2A.1: Starting Sample and Exam-Person-Year Duplicate Share

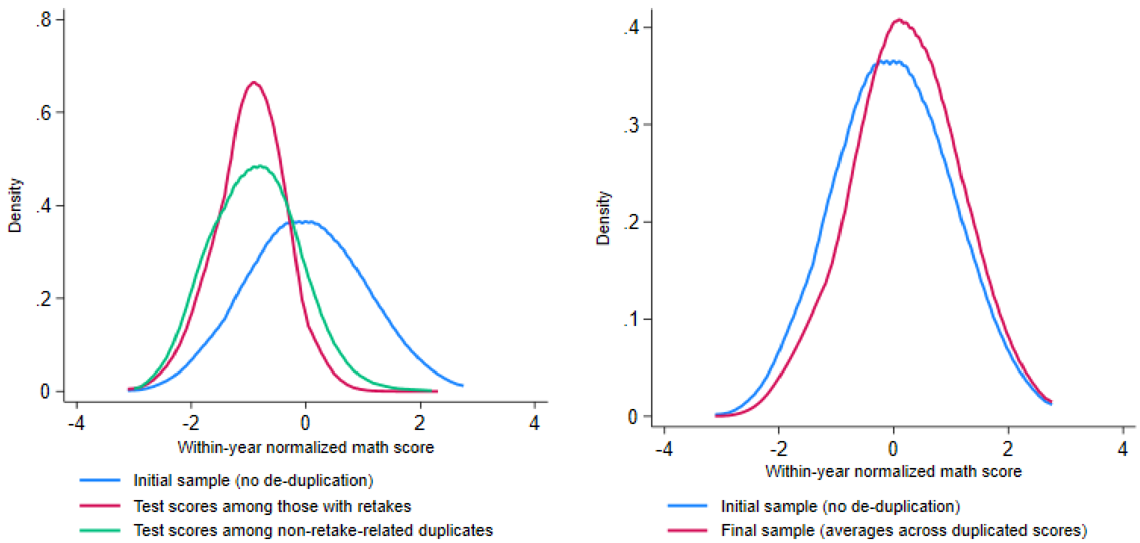
	2009	2010	2011	2012	2013	2014
ACT Math						
Number of students				93,667	95,627	98,121
% with test retake				0%	0%	0%
% other duplicate				0.4%	0.7%	0.5%
Grade 8 Math						
Number of students	104,985	106,157	108,248	109,638	112,738	117,285
% with test retake	21.2%	20.2%	18.9%	0.2%	0%	0%
% other duplicate	1.4%	1.0%	0.8%	0.7%	0.7%	1.7%

Notes: ACT for students in grade 11 starts in 2012, which means I need grade 8 scores starting in 2009 when controlling for past test scores; I present Math exam information because the results from the Reading exam are very similar. The results from 2015 and later are similar to 2012-2014, but they are omitted for brevity. For each the ACT and eighth grade Math test, I present the number of test-takers along with the percentage who are marked as having a retake and the percentage who have a non-retake-related additional test score observation.

A couple of immediate questions emerge from this data pattern. First, is there selection into which students retake exams (i.e., are lower-scoring students more likely to retake the exam)? Second, does taking within-year averages over multiple test scores and removing the duplicates shift the test score distribution? Figure 2A.1 shows that while students that retake tests as well as students with non-retake-related duplicates score lower on average (i.e., there is selection into the group of students with duplicated observations), the impact is small when averaging over duplicates and does not generate cause for concern.

[15] Some of these duplicates are marked as an exam retake, but others are not explained by the data.

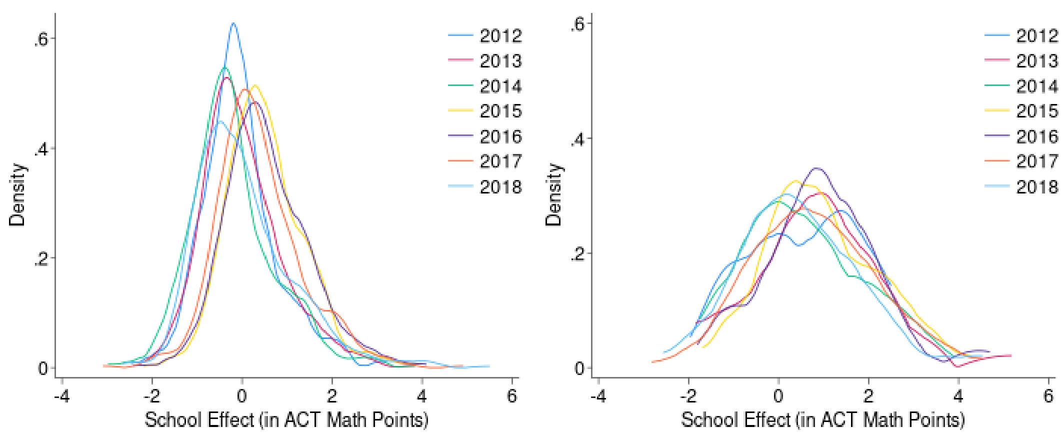
Figure 2A.1: Distributions of 2009 Grade 8 Math Test Scores and Duplicates



Notes: For brevity, I include only 2009's distributions of eighth grade math scores; 2010-2011 are similar. From the panel on the left, it is clear that students with multiple test score observations come from the left side of the distribution. However, from the panel on the right, the change in the overall distribution is small: some of the density from the left half of the initial distribution is removed while the concentration around the mean increases.

2.B Supplementary Tables and Figures

Figure 2B.1: Distributions of School Effects on ACT Math Scores, by Year

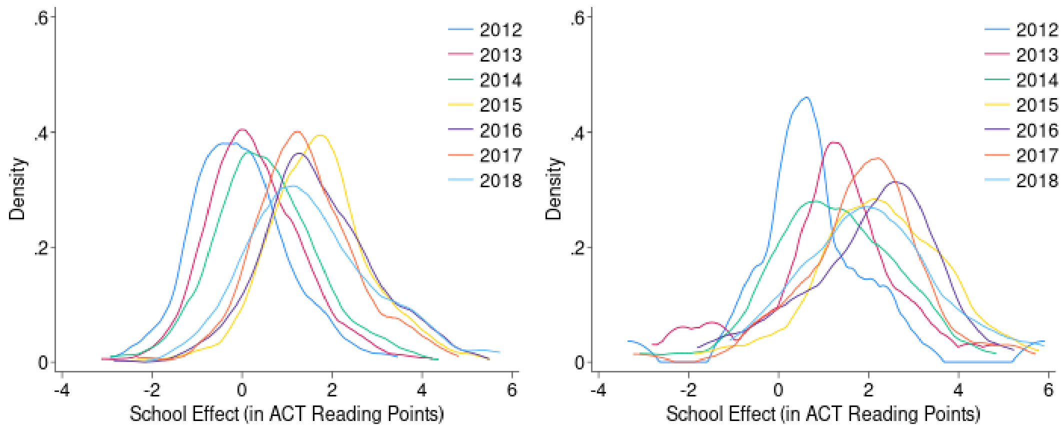


(a) Traditional public schools only

(b) Charter schools only

Notes: Each graph shows distributions of de-meaned school effect estimates; the dependent variable is ACT Math score for robustness of the overall ACT score. The left panel includes only traditional public schools while the right panel includes only charter schools.

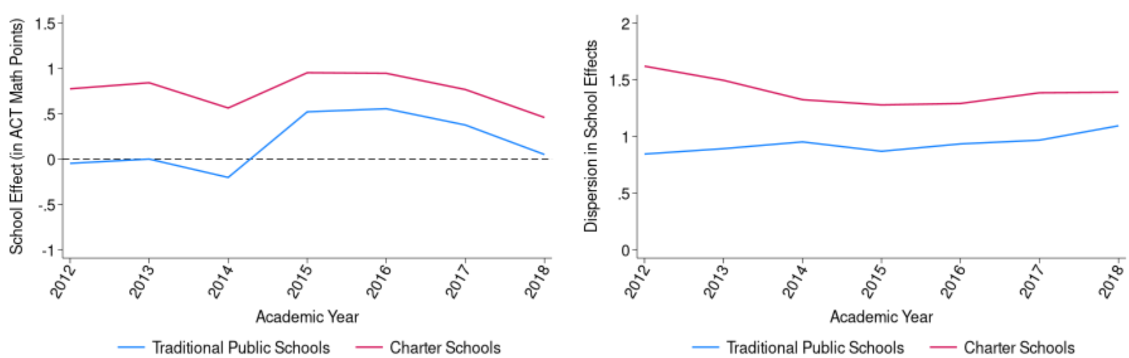
Figure 2B.2: Distributions of School Effects on ACT Reading Scores, by Year



(a) Traditional public schools only (b) Charter schools only

Notes: Each graph shows distributions of de-meaned school effect estimates; the dependent variable is ACT Reading score for robustness of the overall ACT score. The left panel includes only traditional public schools while the right panel includes only charter schools.

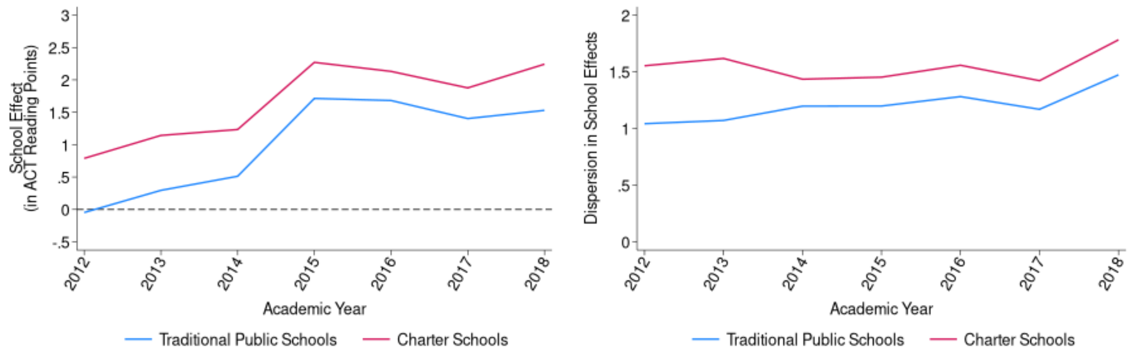
Figure 2B.3: Trends in School Effect Mean and Dispersion Over Time (Using ACT Math Scores)



(a) Mean (b) Dispersion

Notes: Panel (a) shows trends in value-added of the mean traditional public school (blue) and charter school (red); panel (b) shows trends in the dispersion in value-added across schools. The dashed line in panel (a) represents the 2012 mean value-added across all schools. ACT Math scores are used for robustness of the baseline results that use the overall ACT score.

Figure 2B.4: Trends in School Effect Mean and Dispersion Over Time
(Using ACT Reading Scores)

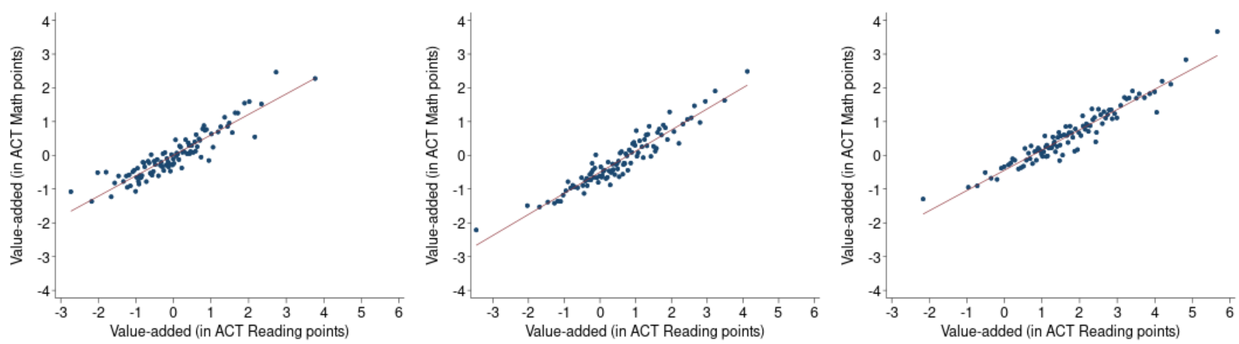


(a) Mean

(b) Dispersion

Notes: Panel (a) shows trends in value-added of the mean traditional public school (blue) and charter school (red); panel (b) shows trends in the dispersion in value-added across schools. The dashed line in panel (a) represents the 2012 mean value-added across all schools. ACT Reading scores are used for robustness of the baseline results that use the overall ACT score.

Figure 2B.5: Correlation Between ACT Math and ACT Reading Value-Added Estimates



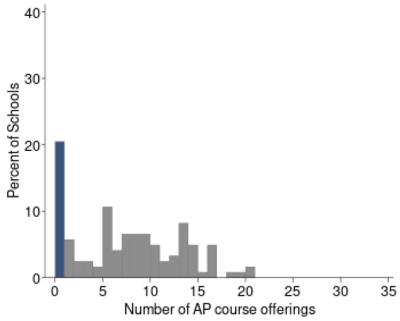
(a) 2012, R-squared = 0.514

(b) 2014, R-squared = 0.588

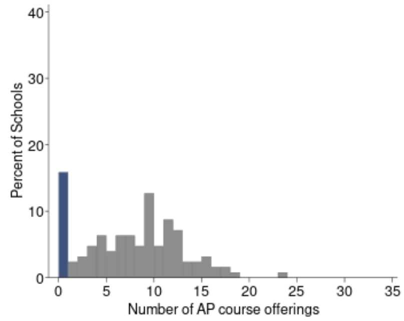
(c) 2016, R-squared = 0.650

Notes: Each graph above is a bin scatter that uses a bin size of 100, meaning each point represents the average among 100 schools. There are positive correlations between the estimates that use ACT Math and ACT Reading scores in each year, indicating that the results are not being driven by the chosen outcome.

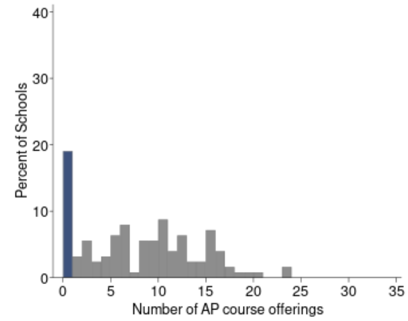
Figure 2B.6: Distributions of Advanced Course Offerings, Conditional on School Quality



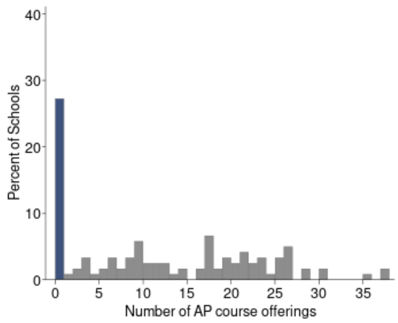
(a) Low quality schools, 2012



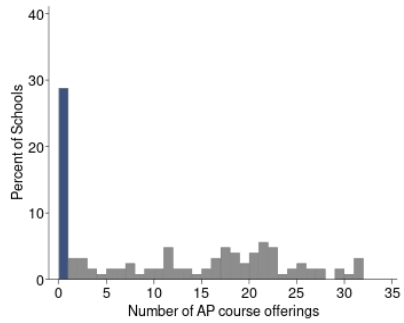
(b) Low quality schools, 2014



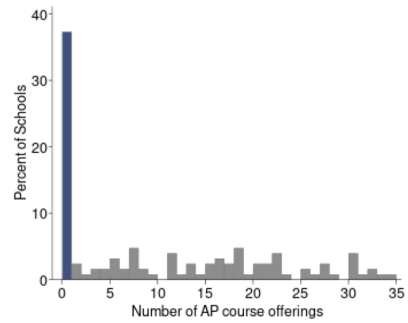
(c) Low quality schools, 2016



(d) High quality schools, 2012



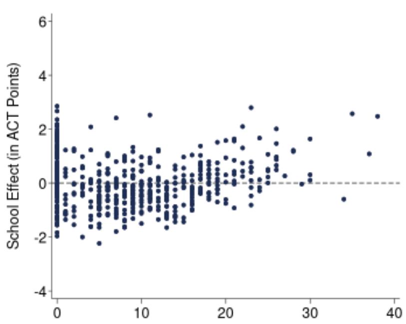
(e) High quality schools, 2014



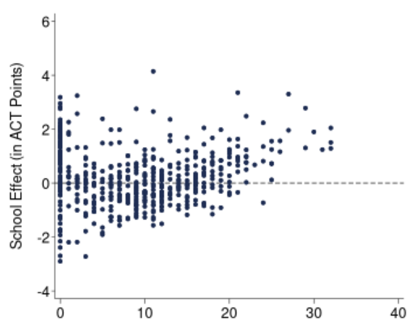
(f) High quality schools, 2016

Notes: Each panel shows a histogram of the number of advanced course offerings across schools, conditional on school quality; the top set of panels include schools in the lowest quartile of the school effect distribution; the bottom set of panels include schools in the highest quartile of the school effect distribution. Only traditional public schools are included. Large variation even when conditioning on school quality indicates that factors other than school quality determine the curriculum.

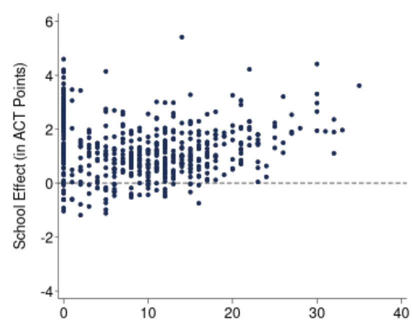
Figure 2B.7: Scatter Plots of Advanced Course Offerings Against School Quality



(a) 2012



(b) 2014



(c) 2016

Notes: Each graph above is a scatter plot of number of advanced course offerings against school quality (using overall ACT scores). The dashed horizontal line represents the quality of the 2012 mean school. For any given value-added estimate, the spread on number of advanced course offerings is incredibly large; moreover, conditional on the number of advanced course offerings, there is variation in school quality.

Table 2B.1: Decomposition of High School Quality Looking at the Extensive Margin of Advanced Course Offerings (Equation (2.7))

	Unweighted		Weighted
	(1)	(2)	(3)
Curriculum:			
AP Math course dummy	0.018	0.059	0.083
	(0.074)	(0.074)	(0.054)
AP Science course dummy	0.009	0.033	0.062
	(0.065)	(0.065)	(0.054)
AP Art course dummy	0.045	0.022	0.044
	(0.041)	(0.034)	(0.035)
AP History course dummy	−0.052	−0.061	−0.024
	(0.081)	(0.071)	(0.066)
AP English course dummy	−0.208	0.025	−0.040
	(0.083)	(0.071)	(0.072)
AP Language course dummy	0.152	0.025	0.028
	(0.054)	(0.045)	(0.038)
Student-teacher ratio	0.011	−0.009	−0.006
	(0.006)	(0.015)	(0.016)
Aggregate teacher characteristics:			
Percent with a master's degree or higher	0.049	−0.026	−0.016
	(0.009)	(0.010)	(0.009)
Average years of experience	0.027	0.024	0.023
	(0.010)	(0.015)	(0.013)
Student body composition:			
Average grade 8 Math score	−0.884	−0.836	−1.095
	(0.189)	(0.314)	(0.275)
Average grade 8 Reading score	1.555	1.194	0.617
	(0.207)	(0.398)	(0.363)
Percent qualifying for subsidized lunch	−0.009	0.004	0.004
	(0.002)	(0.002)	(0.002)
Percent male	0.008	0.002	0.001
	(0.004)	(0.008)	(0.008)
Percent white	−0.013	0.009	0.011
	(0.002)	(0.008)	(0.007)
Principal fixed effects		Y	Y
School fixed effects		Y	Y
Academic year fixed effects		Y	Y
R-squared	0.394	0.842	0.879
School-year observations	3332	3112	3112

Notes: The dependent variable is value-added estimated with overall ACT scores. Weighted specifications weight by total school enrollment; standard errors are clustered at the school level. Only traditional public schools are included in the analysis.

2.C Potential Endogeneity of Advanced Course Offerings

Advanced course offerings may fluctuate from year to year due to factors like the student body composition and student demand. Moreover, principals or other administrators likely play a role in the determination of which courses a school offers. In the main specification, I control for student characteristics and include a principal fixed effect, which should help in isolating the exogenous variation tied to advanced course offerings; one could further employ an instrumental variable strategy around a time-varying component of the 2014 North Carolina Advanced Placement Partnership to *only* utilize exogenous variation in advanced course offerings. Recall that as part of the NCAPP, some of the low-performing school districts are labeled by the State Board of Education as "targets" each year and given prioritized assistance from the NC DPI, particularly in the development of advanced curricula. Letting D_1 be the group of "target" school districts at some point (D_0 contains control districts) and τ_{dt} be the time at which district $d \in D_1$ became a "target" district, one could construct the following time-dependent indicator:

$$\text{NCAPP\_target}_{dt} = \begin{cases} 0 & \text{if } d \notin D_1, \text{ or } d \in D_1 \text{ and } t < \tau_{dt} \\ 1 & \text{if } d \in D_1 \text{ and } t \geqslant \tau_{dt} \end{cases} \tag{2.12}$$

This indicator could be used as an instrument for a curriculum index C_{jt} that encompasses some aggregate of a school's advanced course offerings. More specifically, to estimate the subgroup-specific effect of the number of advanced course offerings on student achievement, the specification in Equation (2.8) could be adapted as follows:

$$\begin{aligned} \text{Stage 2:} \quad Y_{ijst} = {} & \delta^{IV}_{0,k} + \delta^{IV}_{S_k}[\mathbb{1}(i \in S_k) \times \hat{C}_{jt}] + \delta^{IV}_{S'_k}[\mathbb{1}(i \in S'_k) \times \hat{C}_{jt}] \\ & + X'_{it}\,\delta^{IV}_{3,k} + \rho_p + \eta_{js} + \omega_t + \varepsilon_{ipjst} \quad \text{for } k = 1,2,3,4 \end{aligned} \tag{2.13}$$

$$\text{Stage 1:} \quad C_{jt} = \gamma_0 + \gamma_1\,\text{NCAPP\_target}_{dt} + \bar{X}'_{jt}\,\gamma_2 + \zeta_d + \omega_t + u_{jdt} \tag{2.14}$$

This specification could address initial concerns of endogeneity in the main specification, though a concern is null results across the board if the NCAPP does not generate enough variation in advanced course offerings.

3 Equity Concerns and the Impact of One-to-One Technology Initiatives on Student Achievement

3.1 Introduction

Technology has recently pervaded the educational atmosphere, contributing to evolving teaching methods and altered student outcomes. Even prior to the COVID-19 pandemic, more and more schools began experimenting with providing each student with their own computer,[1] a policy that goes by many informal names but is commonly referred to as a one-to-one technology initiative (Penuel, 2006; Escueta et al., 2020). In Wisconsin, this technological shift can be traced back to the June 2010 adoption of the Common Core State Standards,[2] which place emphasis on the integration of technology in academic curricula.

This paper investigates how one-to-one technology initiatives impact students' English Language Arts (ELA) and Math test scores as well as achievement gaps across socioeconomic subgroups.[3] Schools implement one-to-one initiatives hoping to provide students with another tool that can improve transferable skills and better prepare them for the transition from class to career; Kozma (2003) finds that these initiatives are successful in improving students' information and technology skills, which are important in preparing students to move from class to career. In addition, while Wisconsin's one-to-one initiatives rarely operate under the promise of improving student achievement, they aim to improve learning and have the potential to directly impact students' test scores. With documentation that computers quickly become less of a learning aid and more of an integral piece of pedagogy during a one-to-one initiative (Means et al., 2001; Schofield and Davidson, 2002; Machin et al., 2007), students are all the more likely to see impacts on their achievement. Furthermore, with heterogeneity in students' access to internet and other resources at home, these initiatives might impact students from different socioeconomic subgroups differently.

In this paper I use data from two sources: (1) student-level public school records from the Wisconsin Department of Instruction (DPI), and (2) computer distribution plans under one-to-one initiatives at some of Wisconsin's largest public school districts. The combination of these datasets allows for observation of students' test scores, demographics, and implementation of one-to-one initiatives. To answer my research question, I use a two-way fixed effects design based on Goodman-Bacon (2021), which allows for heterogeneity in the timing of computer distribution; I subsequently extend the model to allow for dynamic effects. Since I know *when* a student receives a computer but not *how* the computer is used, this is an intent-to-treat analysis.

[1]The American Community Survey (ACS), which collects data on computer ownership and internet subscriptions, defines computers to be any desktop, laptop, tablet, or smartphone.

[2]In 2009, Common Core was developed as a set of K-12 academic standards in English Language Arts and Mathematics that focuses on nationwide consistency in student learning. Adoption occurs at the state level; as of 2021, there are only nine states that have not adopted these standards.

[3]For simplicity, this paper will deal with two subgroups: students from high-income households and students from low-income households, as determined by subsidized lunch eligibility.

I find that one-to-one initiatives widen ELA achievement gaps by 0.03-0.04 standard deviations (SD), driven by statistically significant negative effects on students from low-income households. With ELA achievement gaps hovering around 0.75 SD, this would indicate that one-to-one initiatives widen the ELA gap by around 5%. These effects on ELA test scores may be due to a combination of mechanisms, including technological dependence of ELA curricula with the adoption of one-to-one initiatives (Means et al., 2001) and limitations of internet and other necessary at-home resources for students from low-income households (Martin, 2021), which creates a pattern termed the "digital divide" (Bulman and Fairlie, 2016; Escueta et al., 2020). With regard to Math test scores, I find that one-to-one initiatives have a null effect on Math achievement gaps as a result of slightly negative effects on Math test scores across the board (0.01-0.02 SD). One mechanism that could explain these effects on Math test scores is that the computers are a distraction in a subject that is less dependent on technology (Belo et al., 2014; Patterson, 2018).

This paper fits into a broad literature that has focused on two aspects of technology in education: information and communication technology (ICT) and computer assisted instruction (CAI). ICT covers any computer hardware or internet connection while CAI relates to specific instruction that is presented on a computer (Bulman and Fairlie, 2016). In the context of my research question, it is likely that as schools began providing computers to students (ICT), they also began using specific software to carry out lesson plans or platforms on which students complete homework assignments (CAI).[4] Because I do not observe specifically how each computer is used on a student-to-student basis - and requirements likely vary greatly by teacher, grade level, and academic year - my paper fits into the ICT branch of literature with an intent-to-treat study.[5]

In developing countries, the most prominent one-to-one initiative is the One Laptop per Child (OLPC) program, which aims to improve learning among disadvantaged children through laptop distribution. Several papers study the large-scale implementation of OLPC in Peru, finding positive effects on cognitive skills and laptop proficiency, but null effects on test scores (Cristia et al., 2014; Beuermann et al., 2015; Cristia et al., 2017). Limited educational gains of one-to-one initiatives are also found in Colombia (Barrera-Osorio and Linden, 2009), Turkey (Aypay, 2010; Sweet and Meates, 2004), Israel (Angrist and Lavy, 2002), and the United States (Fairlie and Robinson, 2013). While null effects of technology on educational outcomes are most commonly found in past literature, there are mixed results due to papers that find positive effects (e.g., Machin et al., 2007; Grimes and Warschauer, 2008; Suhr et al., 2010; Bulut and Delen, 2011) and papers that find negative effects (e.g., Papanastasiou et al., 2003). Much of the past literature has relied on small samples that make it difficult to make definitive conclusions, a result of high costs associated with randomized control trials often utilized in studies of one-to-one initiatives. This paper contributes to the broader literature on one-to-one initiatives with a substantially larger sample.

In addition, past literature on one-to-one initiatives rarely investigate heterogeneous effects on

[4]One prevalent example of CAI is the web-based learning management system Canvas.

[5]There is now a vast literature studying the effects of virtual learning and school closures in the context of the COVID-19 pandemic, but these should be considered a very different treatment from a one-to-one technology initiative that occurred prior to the pandemic.

students from different subgroups, which is where this paper's biggest contribution lies. Kim and Chang (2010) is one of the few exceptions, though they focus on achievement gaps based on English proficiency rather than socioeconomic status. With a well-documented discrepancy in internet and other resources at home between students from low-income and high-income households (Martin, 2021), one-to-one initiatives are more closely related to income-based achievement gaps. So while there is an overlap between students that are not English proficient and students from low-income households, studying heterogeneous effects on students from low-income and high-income households may yield more implications for policy.

A final contribution of this paper is a more rigorous argument for causation than previous studies, with a formal set of identifying assumptions underlying the two-way fixed effects design. While much of the past literature has put forth descriptive evidence of a correlation between one-to-one initiatives and students' educational outcomes, there are only a few papers that argue for a causal impact of one-to-one initiatives on students' test scores (e.g., Angrist and Lavy, 2002; Machin et al., 2007; Barrera-Osorio and Linden, 2009). The empirical strategy of this paper is most comparable to the difference-in-differences approach of Barrera-Osorio and Linden (2009); the other two papers use instrumental variable strategies. This is also one of the first papers to allow for dynamic effects of one-to-one initiatives with an event study.

The remainder of this paper is organized as follows. Section 3.2 discusses my primary sources of data and presents descriptive statistics on the analysis sample. Section 3.3 outlines my empirical strategy, which includes a two-way fixed effects approach followed by an event study design. Section 3.4 presents the main results while Section 3.5 lays the groundwork for a cost-benefit analysis and discusses longer-term impacts of one-to-one initiatives in the context of the COVID-19 pandemic. Finally, Section 3.6 concludes.

3.2 Data and Descriptive Statistics

Data

In this paper I use two sources of data: (1) student-level public school records from the Wisconsin Department of Public Instruction (DPI), and (2) computer distribution schemes under one-to-one initiatives. The school records date back to the 2005-06 academic year and run through the 2020-21 academic year.[6] My primary outcome is standardized test scores; I also observe several demographic characteristics, including student race and ethnicity (condensed into one variable in the data), sex, subsidized lunch eligibility, English language proficiency, and special education enrollment.

In terms of one-to-one technology initiatives, I collected computer roll-out plans from administrators at 10 of Wisconsin's largest public school districts.[7] For these districts - which include (1)

[6]For simplicity, I will refer to an academic year by the fall in which it starts; for example, 2005 represents the 2005-06 academic year. DPI lists the start of an academic year as July and the end of that academic year as the following June.

[7]I have additional data from one school district that provided iPads to its students; however, I only include the school districts that provided Chromebooks in my analysis sample to ensure treatment homogeneity.

Madison Metropolitan, (2) Kenosha, (3) Green Bay Area, (4) Appleton Area, (5) Janesville, (6) Eau Claire Area, (7) Oshkosh Area, (8) Middleton-Cross Plains Area, (9) Sun Prairie Area, and (10) Elmbrook - I observe the academic year in which each grade within each school began providing computers to all students, as well as information regarding computer maintenance, cost, and return policies. Additionally, I have information on how the policy was formulated, which is crucial in demonstrating that the roll-out is tied to factors unrelated to student achievement at baseline.

Background: Subsidized Lunch Eligibility

Subsidized lunch eligibility is determined prior to each academic year primarily based on household income. Children from families with income equal to or below 185% of the federal poverty line (FPL) are eligible for reduced-price lunch at school. The two primary methods for determining eligibility are direct certification and applications submitted prior to the start of the academic year. Direct certification automatically enrolls children in the subsidized lunch program if they participate in FoodShare (SNAP), W-2 cash benefits (TANF), or Food Distribution Programs on Indian Reservations (FDPIR). Students are also categorically eligible if they are determined to be homeless, runaway, migrant, foster, or enrolled in a Head Start program. In this paper, I will refer to students that qualify for subsidized lunch as being from low-income households, and students that do not qualify as being from high-income households. There are certainly students that do not qualify for subsidized lunch but are still from lower-income households, but describing subgroups as "low" and "high" prevents unnecessary wordiness in explaining the results of this paper.

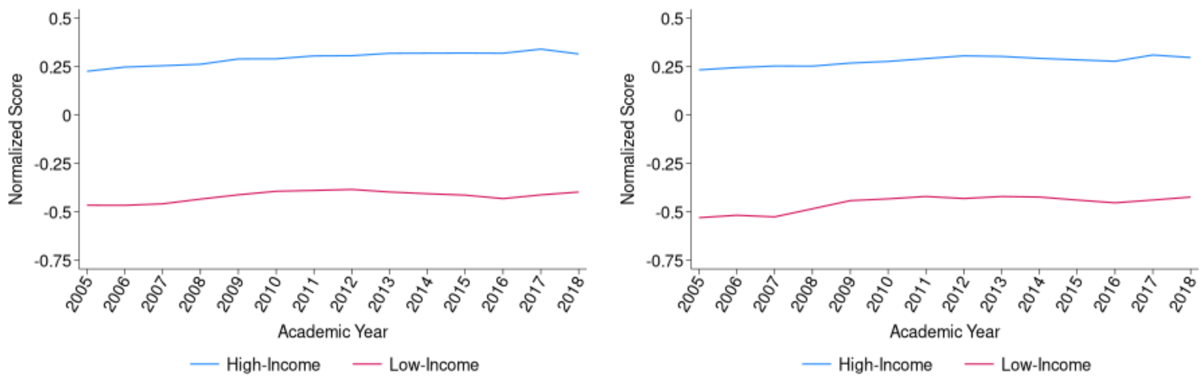
Background: Standardized Exams

The Wisconsin Student Assessment System (WSAS) is the state's suite of standardized examinations given annually to students in the K-12 public school system. Due to updates in legislature and academic standards, many features of the WSAS have changed since 2005, including exam format, timing, grade levels tested, and subject goals. Prior to 2014, the WKCE was administered to students in grades 3-8 and 10 every fall. For students in grades 3-8, the Badger exam replaced the WKCE in 2014, though it was quickly replaced by the Forward exam in 2015; both of these exams took place in the spring. From 2014 onward, the ACT Aspire was administered every spring to students in grades 9-10 while the ACT was administered every spring to grade 11; high school students are not included in this paper's main analysis because there is limited test score data for these students prior to the implementation of Wisconsin's one-to-one initiatives, which began in 2014. Alternate exams are provided for students with significant cognitive disabilities, though they are also not included in my analysis because few students take them and the results are difficult to compare with those from the standard exams. The exam subjects I focus on in this paper are English Language Arts (ELA)[8] and Mathematics, because they are tested most frequently and tied

[8]English Language Arts - which replaced Reading in 2014 due to changing standards - includes several areas of reading and writing, and examines what is broadly referred to as English literacy.

more directly to Common Core.[9]

Figure 3.1: ELA Achievement Gaps Between Students from Low-Income Households and Students from High-Income Households



(a) Elementary schools

(b) Middle schools

Notes: These graphs show longstanding ELA achievement gaps between students from low-income (red line) and high-income households (blue line), which come from subsidized lunch eligibility. The left graph is for students from elementary schools while the right graph is for students from middle schools.

Wisconsin is particularly interesting when analyzing policies related to student equity and achievement gaps. The National Assessment of Educational Progress and Wisconsin DPI find that despite the relatively average overall performance of Wisconsin students on standardized exams, achievement gaps between Black and White students are among the highest in the country, both in ELA and Math (Lee et al., 2007). I find achievement gaps between students that qualify for subsidized lunch and those that do not to be similarly high (around 0.75 SD), as shown in Figure 3.1. This makes Wisconsin an interesting setting to analyze one-to-one technology initiatives.

Background: One-to-One Technology Initiatives

One-to-one initiatives in Wisconsin's public school districts can be traced back to the state's 2010 adoption of the Common Core State Standards, which emphasized integrating technology in academic curricula. The decision to provide computers to students happens at the district level because the purchase of computers is almost entirely dependent on the district budget.[10] Within a district, implementation time varies by school-grade (i.e., every student in grade g at school j receives a computer at the same time). There is no opt-out for students in school-grades that provide computers,[11] and the one-to-one initiatives are designed such that once a student has received a computer, that student will have one in all subsequent years of their K-12 education.[12]

[9]For a detailed timeline of the evolving WSAS, see Figure 3A.1.

[10]Districts typically plan for a four-year computer lifespan and commonly buy the computers outright. Some districts opt to lease computers at a lower total cost instead of buying them.

[11]However, there is nothing to prevent a student from not using a computer if they do not want to.

[12]This feature will be important for the purpose of identification because it means that once a student is "treated" by the policy, they never revert back to the "untreated" state.

Table 3.1: Computer Roll-Out in Wisconsin's One-to-One Initiatives

Implementation Year	Number of School-Grades	Fraction out of Total	In Treatment Group	In Never-Treated Group
2014	6	0.01	Y	
2015	26	0.04	Y	
2016	68	0.11	Y	
2017	88	0.15	Y	
2018	39	0.07	Y	
Later than 2018	370	0.62		Y
Total	597	1	227	370

Notes: 38% of the school-grades in my sample provided computers 2014-2018; these comprise the treatment group. The remaining 62% either planned to provide laptops after 2018 or had no one-to-one initiative; these comprise the never-treated group.

Table 3.1 shows how the computer roll-out progressed in the school districts in my sample. There are a total of 597 school-grades in my sample, 227 school-grades of which compose the treatment group and 370 school-grades of which make up the never-treated group. Very few school-grades implemented in 2014, as this first year is seen as more of a "test phase" among districts during which they learned whether or not the program was helpful or sustainable; the computer distribution begins to really ramp up in 2016.

Descriptive Statistics

Some of the key restrictions that I impose on the data have been outlined above, but I will reiterate them here for the purpose of being clear about the final analysis sample. In terms of the analysis period, I restrict to 2006-2018. On the front end, I am only able to construct lagged test scores starting in 2006 because the data starts in 2005; on the back end, the pandemic disrupted standardized testing after 2018.[13] I also restrict to students in grades 3-8 because (1) elementary school students take their first standardized exam in grade 3, and (2) there is limited test score data for high school students in the years prior to 2014, which is the first year of any one-to-one technology initiative that I observe. And lastly, I am only able to include students from school districts for which I obtained data on their one-to-one initiative.

Table 3.2 presents some descriptive statistics for students in the analysis sample as well as students grades 3-8 statewide. To show that the composition of the sample is relatively stable compared to statewide averages, I present two years of data in Table 3.2, statistics for additional years in Table 3A.1, and trends in Figure 3A.3 in Appendix 3.A. Overall, my sample is more diverse than the population of Wisconsin public school students (fewer White students), though the sample averages are not far off of the state averages. Test scores of students in my sample are falling slightly over time relative to statewide averages, a pattern that may warrant additional investigation. Overall,

[13] The 2019 exam would have occurred in the spring of 2020.

there is evidence that Wisconsin's largest districts are fairly representative of the state, and so with caution one can extend the results of this paper to school districts statewide.

Table 3.2: Descriptive Statistics of Students Grades 3-8 in the Sample and Statewide (2006 and 2018: First and Last Year of the Analysis Window)

	2006 Academic Year		2018 Academic Year	
	Sample (1)	Statewide (2)	Sample (3)	Statewide (4)
Demographic characteristics				
White	0.741	0.768	0.568	0.685
Black	0.099	0.108	0.095	0.091
Hispanic	0.081	0.074	0.183	0.129
Male	0.506	0.512	0.512	0.513
Eligible for subsidized lunch	0.329	0.337	0.471	0.427
Limited English proficient	0.082	0.057	0.117	0.064
Enrolled in special education	0.145	0.133	0.126	0.132
Test scores				
Average grade 5 ELA score	0.088	0	−0.042	0
Average grade 5 Math score	0.110	0	−0.029	0
Average grade 8 ELA score	0.069	0	−0.104	0
Average grade 8 Math score	0.134	0	−0.051	0
Number of unique schools				
Elementary	146	1052	147	843
Middle	44	565	43	823
Number of unique students	41,820	365,230	42,858	365,970

Notes: This table presents descriptive statistics for students in the main analysis sample (columns (1) and (3)) and for students in grades 3-8 statewide (Columns (2) and (4)). The restriction of grades 3-8 comes from (1) the fact that students begin taking standardized exams in grade 3, and (2) I focus on elementary and middle school students in this paper based on the high school standardized exam schedule. Columns (1) and (2) are for the 2006 academic year while Columns (3) and (4) are for 2018; these are the first and last years of my analysis window, respectively. For statistics for additional years, see Appendix 3.A.

An additional concern is that students that received computers (the treatment group) are inherently different from students that did not receive computers (the never-treated group). Table 3.3 shows demographic characteristics and 2013 test scores of each group; 2013 is the last year before any school-grade provided computers to all students so it gives an idea of how things looked prior to the start of any one-to-one initiative. Demographically, the two groups are similar overall; however, in terms of test scores the treatment group has higher test scores at baseline, regardless of school type. This further confirms that districts did not take into account student achievement or student need when formulating their computer roll-out plans: if they had, then students with lower test scores at baseline would have received computers first.

Table 3.3: Descriptive Statistics for Students that Received Computers (Treatment Group) and Students that Never Received Computers (Never-Treated Group)

	Elementary School		Middle School	
	Treatment Group (1)	Never-Treated Group (2)	Treatment Group (3)	Never-Treated Group (4)
Demographic characteristics				
White	0.631	0.620	0.630	0.669
Black	0.122	0.098	0.107	0.104
Hispanic	0.105	0.172	0.133	0.145
Male	0.510	0.511	0.510	0.512
Eligible for subsidized lunch	0.373	0.494	0.435	0.417
Limited English proficient	0.102	0.128	0.111	0.070
Enrolled in special education	0.134	0.135	0.144	0.124
Test scores				
Average 2013 ELA score	0.078	−0.096	-	-
Average 2013 Math score	0.162	−0.059	-	-
Average 2013 ELA score	-	-	0.010	−0.005
Average 2013 Math score	-	-	0.067	0.006
Number of unique students	37,626	90,913	93,065	46,891

Notes: This table presents descriptive statistics for students that received computers at some point in 2006-2018 (the treatment group), and students that never received computers (the never-treated group). Test scores from 2013 are presented because this is the last year prior to the start of any district's one-to-one initiative. More of the middle school students received laptops because the laptop distribution policies tended to start with older students.

3.3 Empirical Strategy

Two-Way Fixed Effects Model

To answer my research questions, I utilize a two-way fixed effects approach, as specified in Goodman-Bacon (2021). As a result of two important features of one-to-one technology initiatives in Wisconsin, my research design boils down to a simple difference-in-differences (DD) with a pre-implementation period and a post-implementation period (Bertrand et al., 2004). First, there are only two groups: (1) students that received computers and (2) students that did not. And second, once a student received a computer, they retain the computer until they graduate from high school (i.e., the treatment never shuts down once it begins). The treatment group consists of students that received computers at some point between 2005 and 2018 while the never-treated group[14] consists of students that had not yet received laptops.[15] To construct the treatment variable, I let $D = 1$ represent the treatment group and $D = 0$ represent the never-treated group. Next, I let $\tau(g, j)$ denote the school year in which grade g in school j began providing students with computers. The treatment T is then a dummy variable that takes the value one for grade g in school j in all years after (and including) $\tau(g, j)$, otherwise T takes the value zero. More formally,

$$T_{gjt} = \begin{cases} 0 & \text{if } D = 0 \\ 0 & \text{if } D = 1, t < \tau(g, j) \\ 1 & \text{if } D = 1, t \geqslant \tau(g, j) \end{cases} \tag{3.1}$$

In a difference-in-differences framework, we can think about this treatment variable as an interaction between an indicator for being in the treatment group and an indicator for whether treatment has occurred, as follows:

$$T_{gjt} = \mathbb{1}[D = 1] \times \mathbb{1}[t \geqslant \tau(g, j)] \tag{3.2}$$

Next, define student subgroups $h \in s$ and $l \in s$ for high-income (students that do not qualify for subsidized lunch) and low-income (students that do qualify for subsidized lunch), respectively; $i \in l$ indicates that student i is from a low-income household. With this in mind, the formal econometric model can be written as follows:

$$Y_{it} = \beta_0 + \beta_h \left[T_{gjt} \times \mathbb{1}[i \in h]\right] + \beta_l \left[T_{gjt} \times \mathbb{1}[i \in l]\right] + X'_{it}\, \beta_X + \gamma_{gjs} + \omega_t + \varepsilon_{it} \tag{3.3}$$

On the left-hand side are students' standardized test scores Y_{it}; I will present results that use Math and ELA test scores. On the right-hand side, I interact the treatment variable with low-income and high-income indicators, which come from subsidized lunch eligibility. This yields a simple and intuitive interpretation of the β_h and β_l coefficients: β_h is the effect of one-to-one initiatives on students from high-income households, and β_l is the effect of one-to-one initiatives on students from low-income households. The difference between these two effects can be interpreted as the effect of one-to-one initiatives on achievement gaps between students from low-income and high-income households. Since I observe implementation of one-to-one initiatives, but not how students use the computers in school or at home, these effects carry intent-to-treat interpretations.

The covariate vector X_{it} controls for lagged test scores and student demographics, including gender, race/ethnicity, attendance rate, and indicators for special education enrollment and limited English proficiency. The most important covariate is lagged test scores, which control for any unobserved inputs to students' test scores. Subsidized lunch status is interacted with a school-grade fixed effect instead of being controlled for in X_{it} to eliminate inherent differences across subgroups

[14]In the context of a DD with variation in treatment timing, it is more appropriate to use the term "never-treated" than "comparison" because the composition of the comparison group changes from year to year.

[15]There are two types of students in the never-treated group: (1) students scheduled to receive computers after 2018, (2) students in schools that hoped to distribute computers but had not yet formulated a roll-out plan. In either case, these students did not receive laptops at any point during the analysis window.

from the subgroup-specific treatment effects I am interested in estimating. In addition, because technology initiatives are decisions made by administrators at the school districts based on flexibility in the district budget, I also include in X_{it} the district-average spending per pupil.

To isolate the treatment effects, I include a school (j)-grade (g)-subgroup (s) fixed effect γ_{gjs} and an academic year fixed effect ω_t. As mentioned, Wisconsin's one-to-one initiatives are implemented at the school-grade level, so the school-grade-subgroup fixed effect will absorb any factors across school-grades that are constant across time as well as purge the main treatment effect parameters of any inherent differences across subgroups. The error term ε_{it} contains any other unobserved variation in test scores. Because I am working with a panel dataset in which students are observed in years leading up to reception of a computer as well as in years after receiving one, I cluster the standard errors at the student level. The model is estimated separately for students in elementary schools and students in middle schools.

Assumptions

I am interested in estimating the effect of one-to-one initiatives on students from high-income households and students from low-income households. In a perfect world, to find the average treatment effect on treated students (ATT), I would compare these students' test scores in the treated state (Y_1) to their test scores in the untreated state (Y_0). In practice, the issue is that I do not observe what treated students' test scores *would have been* had they not been treated (i.e., the counterfactual Y_0 is not observed). In a DD design, we estimate this counterfactual by using a control group of students that did not receive a laptop. Then under the classic DD assumptions, an OLS regression yields unbiased estimates of the ATT on each subgroup of students, as shown below:

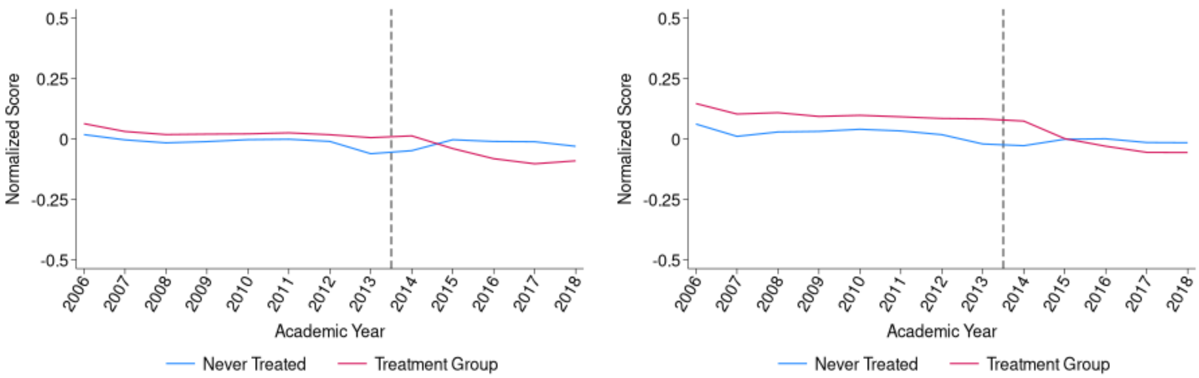
$$ATT_h = E[Y_1 - Y_0 \mid X_{it}, i \in h] = E[Y_1 \mid X_{it}, i \in h] - E[Y_0 \mid X_{it}, i \in h] = \beta_h \tag{3.4}$$

$$ATT_l = E[Y_1 - Y_0 \mid X_{it}, i \in l] = E[Y_1 \mid X_{it}, i \in l] - E[Y_0 \mid X_{it}, i \in l] = \beta_l \tag{3.5}$$

The crucial assumption for identification is that - in the absence of treatment - test scores of the treated group would have trended the same as test scores of the control group, conditional on observables. Researchers often show evidence of parallel pre-trends and argue that these parallel trends would have persisted in the absence of treatment. To this end I show in Figure 3.2 that there are parallel pre-trends in test scores when comparing the treatment group to the never-treated group, conditional on observable characteristics.

With variation in treatment timing, things are more complicated than a comparison between students that received a laptop at some point in the analysis window (2006-2018) and students that did not. Instead, the test scores of students that received laptops in year t are compared to test scores of (1) students that received laptops prior to year t, (2) students that are scheduled to receive laptops after year t, and (3) students that will never receive laptops. In other words, each treatment group is compared to a pool containing the already-treated group, the not-yet-treated group, and the never-treated group.[16] For a clean identification in the DD design I assume that there are no dynamics in the treatment effect, otherwise already-treated students would not be a valid comparison for newly-treated students. This assumption will be relaxed in the next section with an event study.

Figure 3.2: Test Score Trends by Treatment Status (Never Treated and Treated Sometime)



(a) ELA test scores

(b) Math test scores

Notes: These graphs show trends in test scores among students that never received a laptop (blue line) and students that received a laptop from their school 2018 or earlier (red line). The dashed vertical line falls just before 2014 - the year in which the first group of students are given laptops. For simplicity, all students that received laptops are pooled into a single treatment group here; for a more detailed breakdown of the treatment group trend based on the year of laptop receipt, see Appendix 3.A. Elementary and middle schools are pooled here for simplicity, even though they are analyzed separately in the model.

There are three additional criteria that must be satisfied to uncover unbiased effects of one-to-one initiatives. First, the distribution of laptops must be unrelated to students' test scores at baseline. According to the Wisconsin DPI, individual school districts signed one-to-one initiatives into action subsequent to the state's adoption of Common Core in 2010 when they determined revenue[17] to be available for the purchase of computers. This would tie variation in laptop distribution across school districts to variation in district budgets rather than exam-based student need, which is why I control for district spending in the model. Furthermore, distribution of laptops within school districts is based on the curriculum, which often means older students receive laptops first because technology may be more pertinent to their schoolwork.[18] For these key reasons, I have no concern that one-to-one initiatives were influenced by students' test scores.

The second criterion is that the composition of the treatment and control groups are stable over time. Since there is variation in treatment timing, students enter the treatment group (and leave

[16] Figure 3A.4 in Appendix 3.A shows test score trends that break the treatment group down based on year of treatment; these trends are not as smooth as the pooled version due to smaller samples of treated students.

[17] Annual revenue in Wisconsin's public school districts is derived from state and federal aid, as well as property taxes; administrators meet on a regular basis to determine how to spend the revenue.

[18] This type of roll-out that started in high schools, moved to middle schools, and ended in elementary schools also ensures that students retain their laptops even after moving from elementary to middle school or middle to high school. This means that students never leave the treatment group once they enter it.

the control group) in different years; it is important that the control group remains an adequate comparison for the treatment group throughout the analysis window. Table 3.4 shows that the demographic composition of the treatment and control groups are fairly stable over time, although the treatment group grows closer in composition to that of the control group over time. Since the size of the treatment group is growing over time, it is unsurprising that it becomes more representative of the population from year to year.

Third, I assume that there are no spillover effects; in other words, students that are not provided with a laptop are not impacted by their peers that do receive a laptop. The way in which school districts distribute laptops should minimize any concern of this: every student in a particular school-grade receives a laptop at the same time. Under the assumption that the majority of a student's network is in the same grade as them, there shouldn't be any spillover effects of Wisconsin's one-to-one technology initiatives.

Table 3.4: Composition of the Treatment and Control Groups Across Years

	2013 Academic Year		2016 Academic Year		2018 Academic Year	
	Treatment (1)	Control (2)	Treatment (3)	Control (4)	Treatment (5)	Control (6)
Demographics						
White	-	0.628	0.597	0.585	0.560	0.576
Black	-	0.092	0.102	0.089	0.102	0.087
Hispanic	-	0.158	0.132	0.190	0.168	0.199
Male	-	0.514	0.521	0.509	0.511	0.512
Elig. for subsidized lunch	-	0.460	0.353	0.475	0.454	0.489
Limited English prof.	-	0.091	0.084	0.101	0.119	0.115
Enrolled in special ed.	-	0.128	0.122	0.128	0.129	0.123
Test scores						
Avg grade 5 ELA score	-	−0.045	0.078	−0.118	0.054	−0.084
Avg grade 5 Math score	-	0.006	0.184	−0.115	0.109	−0.090
Avg grade 8 ELA score	-	0.005	−0.169	−0.091	−0.158	0.000
Avg grade 8 Math score	-	0.045	−0.065	−0.060	−0.124	0.088
Number of unique students	-	43,012	9,809	32,459	21,533	21,325

Notes: This table presents descriptive statistics for the treatment and control groups for three years of the data. The 2013 academic year occurs prior to the start of any district's laptop roll-out, so it provides statistics on the full sample. In this table, the treatment and control groups are changing over time due to variation in treatment timing: the treatment group in year t includes students that receive laptops in t or earlier while the control group in year t includes students that hadn't yet received a laptop.

Event Study

In this section I allow for dynamic treatment effects with an event study design, which is particularly important because of the staggered nature of the computer roll-out across schools-grades. In line with past literature (Abraham and Sun, 2021; Borusyak et al., 2024), I include in the model from Equation (3.3) placebo policies for each year before and after the true treatment. More formally, I define treatment variables for each year as follows:

$$T_{gj,t+\kappa} = \begin{cases} 0 & \text{if } D = 0 \\ 0 & \text{if } D = 1, t \neq \tau(g,j) + \kappa \\ 1 & \text{if } D = 1, t = \tau(g,j) + \kappa \end{cases} \tag{3.6}$$

I am no longer interested in estimating the policy effect when moving from the untreated state to the treated state; instead, I want single year effect when moving from school year $t + \kappa - 1$ to school year $t + \kappa$, the time at which $T_{gj,t+\kappa}$ flips from 0 to 1. The formal econometric model for this event study is written as follows:

$$\begin{aligned} Y_{it} = \beta_0 &+ \sum_{\kappa=-4,\kappa\neq-1}^{\kappa=3} \Big[\beta_{\kappa,h} \left[T_{gj,t+\kappa} \times \mathbb{1}[i \in h]\right]\Big] \\ &+ \sum_{\kappa=-4,\kappa\neq-1}^{\kappa=3} \Big[\beta_{\kappa,l} \left[T_{gj,t+\kappa} \times \mathbb{1}[i \in l]\right]\Big] + X'_{it}\,\beta_X + \gamma_{gjs} + \omega_t + \varepsilon_{it} \end{aligned} \tag{3.7}$$

The first summation in Equation (3.7) encompasses an interaction between the yearly treatment variable and a high-income indicator (student does not qualify for subsidized lunch); the second summation encompasses an interaction between the treatment variable and a low-income indicator. $\beta_{\kappa,h}$ represents the effect of a one-to-one initiative implemented in year t on students from high-income households during year $t + \kappa$; $\beta_{\kappa,h}$ has an analogous interpretation for students from low-income households.

The covariate vector X_{it} is identical to the one in Equation (3.3), and contains lagged test scores, gender, race/ethnicity, attendance rate, indicators for special education enrollment and limited English proficiency, and district-average spending per pupil. I again include a school-grade-subgroup fixed effect γ_{gjs} and an academic year fixed effect ω_t. Standard errors are clustered at the student level, and the model is estimated separately for students in elementary schools and students in middle schools.

3.4 Main Results

Two-Way Fixed Effects

In Table 3.5 I provide the main results from the two-way fixed effects model specified in Equation (3.3), which I estimate separately for elementary and middle schools; Columns (2) and (4) include all covariates exactly as written in Equation (3.3). When using ELA test scores as the dependent variable, I find that one-to-one initiatives widen ELA achievement gaps by 0.03-0.04 standard deviations (SD), both in elementary schools and in middle schools. With slightly positive effects on

Table 3.5: Effects of One-to-One Initiatives on Test Scores and Achievement Gaps (Equation (3.3))

	Elementary School		Middle School	
	(1)	(2)	(3)	(4)
	Outcome: ELA test scores			
High-income	0.015	0.005	0.018	0.008
	(0.008)	(0.008)	(0.005)	(0.005)
Low-income	−0.028	−0.032	−0.014	−0.014
	(0.010)	(0.010)	(0.006)	(0.006)
Achievement gap	0.042	0.037	0.032	0.022
	(0.011)	(0.011)	(0.006)	(0.006)
R-squared	0.660	0.672	0.684	0.696
Student-year observations	228,423	227,290	326,889	326,771
	Outcome: Math test scores			
High-income	−0.006	−0.011	−0.014	−0.024
	(0.007)	(0.007)	(0.005)	(0.005)
Low-income	−0.009	−0.010	−0.012	−0.017
	(0.011)	(0.011)	(0.007)	(0.007)
Achievement gap	0.004	−0.001	−0.002	−0.007
	(0.012)	(0.012)	(0.007)	(0.007)
R-squared	0.646	0.658	0.700	0.711
Student-year observations	230,213	229,068	328,209	328,081
School-grade-subgroup fixed effects	Y	Y	Y	Y
Academic year fixed effects	Y	Y	Y	Y
Baseline covariates	Y	Y	Y	Y
Additional covariates		Y		Y

Notes: This table presents results from the two-way fixed effect regression specified in Equation (3.3) with ELA and Math test scores on the left-hand side. All regressions include school-grade fixed effects and academic year fixed effects. In addition, all regressions include baseline covariates of lagged test scores and an indicator for subsidized lunch eligibility (which is necessary to remove the main income effect from the subgroup-specific effect of one-to-one initiatives). Additional covariates include student gender, race/ethnicity, special education indicator, limited English proficiency indicator, attendance rate, and average district spending per pupil. The effect on achievement gaps is calculated by subtracting the estimated effect on students from low-income households from the estimated effect on students from high-income households. Standard errors are clustered at the student level. Robustness to alternative fixed effects and covariates as well as heterogeneous effects of one-to-one initiatives by grade level are included in Appendix 3.A.

students from high-income households, the negative effects on ELA achievement gaps are driven by statistically significant negative effects on students from low-income households. For context, ELA achievement gaps hover around 0.75 SD prior to the implementation of one-to-one initiatives, as shown earlier in Figure 3.1; this would indicate that one-to-one initiatives widen the gap by around 5%. When using Math test scores as the dependent variable, I find slightly negative effects of one-to-one initiatives on both student subgroups (0.01-0.02 SD), resulting in a null effect on Math achievement gaps. The most interesting finding regarding Math test scores is that policy effects grow more negative when moving from elementary schools to middle schools.

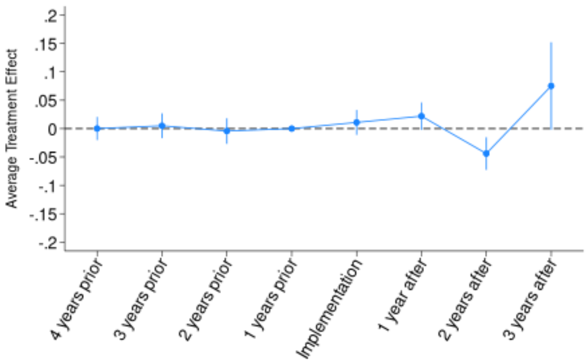
Though untestable - given that I do not observe how students use the computers in school or at home - there are some likely mechanisms that explain these findings. With regard to ELA test scores, effects that are statistically different from zero pretty much across the board (positive for students from high-income households and negative for students from low-income households) may be explained in part by a technological dependence of ELA curricula as a result of one-to-one initiatives. In addition, students from low-income households may not be able to fully utilize the computers due to limitations of internet and other necessary resources at home. Technology that students from low-income households are required to use for school but unable to fully use at home would explain negative effects of one-to-one initiatives on these students' test scores as well as negative effects on achievement gaps. Students from high-income households likely have fewer obstacles to access, and may already have their own computers at home. These mechanisms are supported by results from the 2018 American Community Survey (Martin, 2021), which finds that over 97% of households making over $150,000 per year had an internet subscription while only 65% of households making under $25,000 did. This income-induced gap in access to reliable internet contributed to what is known as a "digital divide" (Bulman and Fairlie, 2016).

With regard to Math test scores, there is likely more separation between curricula and technology, especially in elementary and middle schools where students are still learning basic arithmetic and algebra. However, just as schools cannot force students to use computers for schoolwork, they have little power to prevent students from using the computers to distract themselves from schoolwork. The negative effects of one-to-one initiatives on Math test scores across the board may be due to the computers serving as a distraction rather than a learning aid; there has been some past literature documenting technology as a distraction in school (Belo et al., 2014; Patterson, 2018).

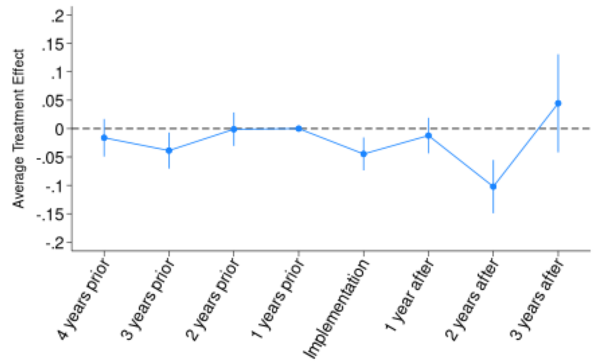
Event Study

This section presents results from the event study design specified in Equation (3.7), which allows for dynamic treatment effects of one-to-one initiatives. Among students from high-income households in elementary schools, I find that there are null effects on ELA and Math test scores in the short-term (one year or less after implementation); these effects dip two years after implementation then become positive in the final year. Findings are similar among students from low-income households in elementary schools, but with slightly negative effects in the first three years followed by a positive effect in the final year. Null or negative effects in the short-term followed by a longer-term positive effect may be the result of an adjustment period in which these younger students grew accustomed to their computers and learned how to use them for their studies.

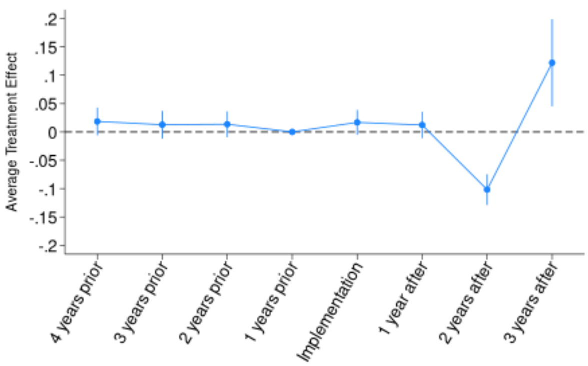
Figure 3.3: Dynamic Effects on Elementary School Students' Test Scores



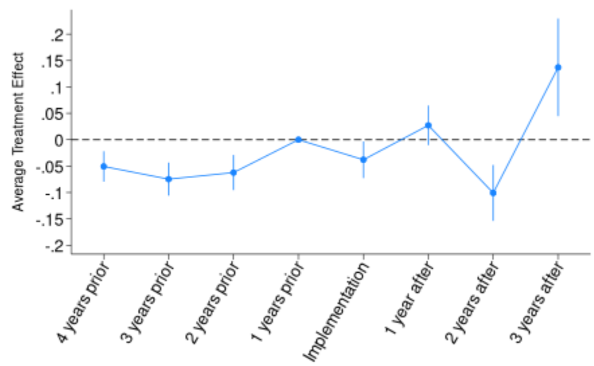
(a) ELA test scores, high-income



(b) ELA test scores, low-income



(c) Math test scores, high-income

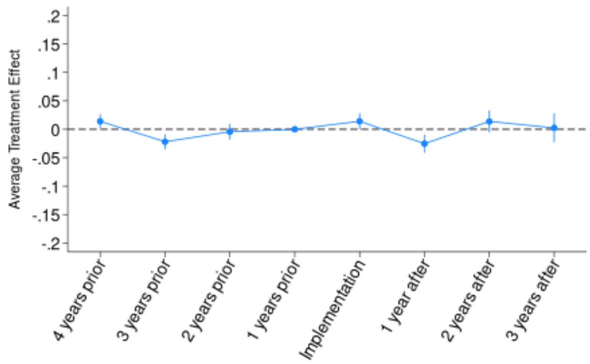


(d) Math test scores, low-income

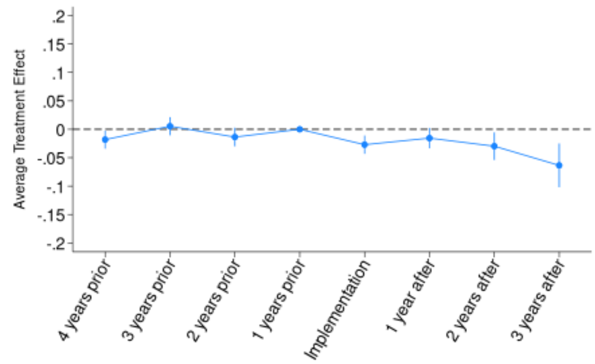
Notes: These graphs show estimated dynamic effects of one-to-one initiatives on elementary school students (see Equation (3.7)). The top set of graphs use ELA test scores as the outcome while the bottom set use Math test scores. The left set of graphs show effects on students from high-income households (those that do not qualify for subsidized lunch) while the right set show effects on students from low-income households. Vertical lines represent 95% confidence intervals.

Among students from high-income households in middle schools, I find effects on ELA test scores that hover around zero for the duration of the post-implementation period; the negative effects on Math test scores from the difference-in-differences specification can be decomposed into negative effects in the short-term followed by a positive effect in the longer-term. Among students from low-income households in middle schools, short-term negative effects on ELA test scores persist into the longer-term and even become more negative over time. Findings are similar when looking at Math test scores, though they are less pronounced and not statistically significant. These results support the theory that students from low-income households lack at-home resources such as reliable internet necessary to succeed in a technology-dependent curriculum.

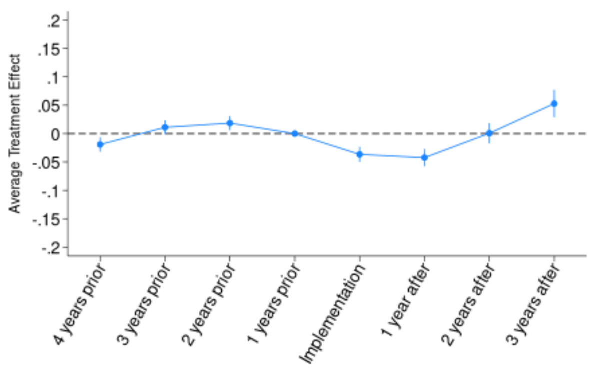
Figure 3.4: Dynamic Effects on Middle School Students' Test Scores



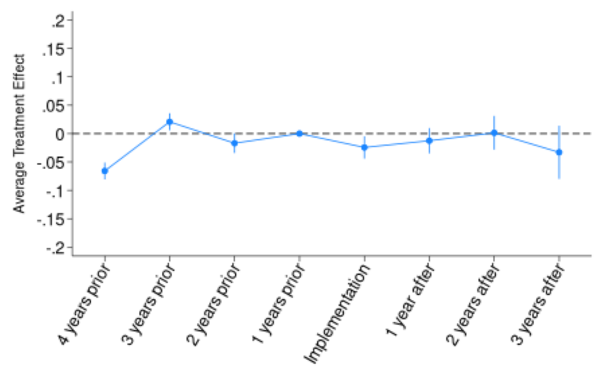
(a) ELA test scores, high-income



(b) ELA test scores, low-income



(c) Math test scores, high-income

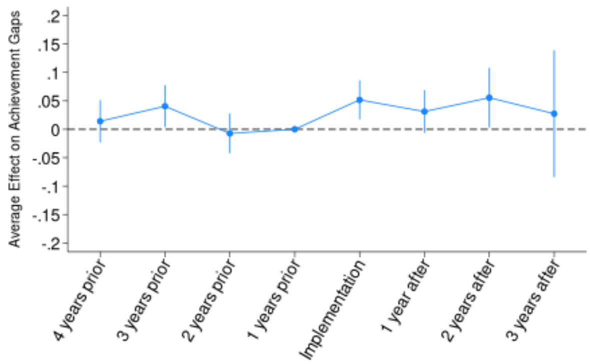


(d) Math test scores, low-income

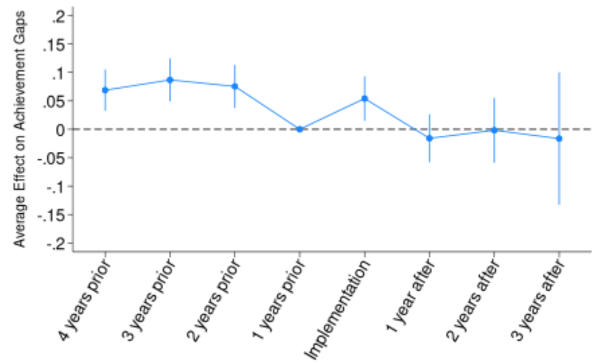
Notes: These graphs show estimated dynamic effects of one-to-one initiatives on middle school students (see Equation (3.7)). The top set of graphs use ELA test scores as the outcome while the bottom set use Math test scores. The left set of graphs show effects on students from high-income households (those that do not qualify for subsidized lunch) while the right set show effects on students from low-income households. Vertical lines represent 95% confidence intervals.

Figure 3.5 shows dynamic effects of one-to-one technology initiatives on academic achievement gaps between students from low-income and high-income households. In elementary schools I find slightly positive impacts on achievement gaps in the short-term, which recede in the longer-term. In middle schools I find null or even slightly negative impacts on achievement gaps in the short-term, which become positive in the longer-term. The patterns in elementary schools suggest that negative short-term effects on students from low-income households generated by an adjustment to the computers go away over time. However, widening achievement gaps over time in middle schools suggest that either a lack of internet and other at-home resources among students from low-income households or higher susceptibility to distractions from schoolwork might have negative impacts on these students that persist into the long-term.

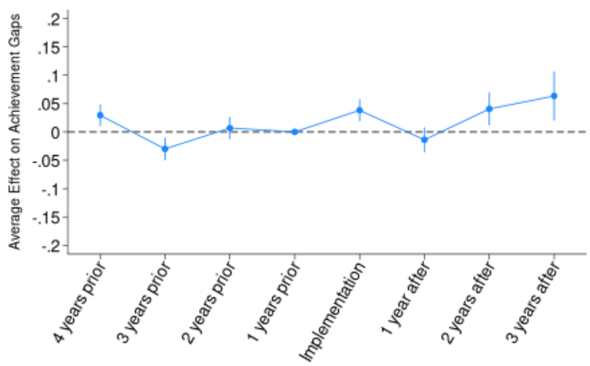
Figure 3.5: Dynamic Effects on Income-Based Achievement Gaps



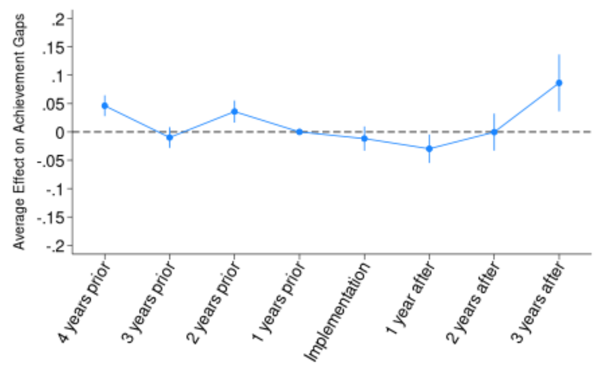
(a) ELA test scores, elementary



(b) Math test scores, elementary



(c) ELA test scores, middle



(d) Math test scores, middle

Notes: These graphs show estimated dynamic effects on achievement gaps between students from low-income and high-income households. The top set of graphs are for the sample of elementary school students while the bottom set are for middle school students. The left set of graphs use ELA test scores as the outcome while the right set use Math test scores. Vertical lines represent 95% confidence intervals.

Since I am using a subsidized lunch indicator to place students in the low-income and high-income groups, the achievement gaps I measure in this paper are understated compared to what they would look like if I we were to split at the 50th percentile. As a result, the impacts I estimate on achievement gaps are likely also lower bounds. In other words, a cleaner split between students from low-income and high-income households would likely yield more heterogeneity in at-home resources, larger differences in treatment effects, and more bite to the mechanisms described above. Another concern with the event study is that the composition of the treatment group is changing from year to year. The latest period $\tau + 3$ will only include school-grades that distributed computers in 2014 or 2015, since the data ends in 2018. In Appendix 3.A, I include a set of figures that only include school-grades that are present in every period.

3.5 Discussion

Cost-Benefit Analysis

Much of the evidence presented thus far indicates that there are null or modest effects of one-to-one initiatives. Are they worth it? Here I aim to answer that question with a simple cost-benefit analysis; with primarily null effects, this analysis will use the most optimistic end of the 95% confidence intervals surrounding estimates in Columns (2) and (4) from Table 3.5. In terms of cost, many of the larger districts in Wisconsin buy computers for their students outright for roughly \$400 per computer over a four-year lifespan, resulting in a cost of \$100 per student per year to the school districts.[19] Districts provide computers for students grades 3-8 (6 years total), so the maximum cost for any single student is \$600.

I now convert test score impacts of one-to-one initiatives into a dollar amount with the help of some past literature. Chetty et al. (2014a) find that a 1 SD improvement in teacher value-added boosts Math test scores by 0.14 SD while Chetty et al. (2014b) tie this benefit to a \$7000 increase in present-value lifetime earnings.[20] Following Jackson and Makarin (2018), I assume causation between the change in test scores and the change in lifetime earnings. With a mean age in my sample also around 12 years, I use the \$7000 figure calculated in Chetty et al. (2014b), but inflate to 2018 dollars in order to more accurately compare to the computer cost. Altogether, I find that a 0.14 SD increase in Math test scores results in a discounted lifetime earnings increase of \$8000.

From Table 3.5, one-to-one initiatives have an average impact of -0.006 SD on Math test scores of students from high-income households and an average impact of 0.014 SD on Math test scores of students from low-income households.[21] This would suggest that students from high-income households see a $\frac{0.006}{0.14} \times \$8000 \approx \$343$ decrease in discounted lifetime earnings; students from low-income households see a $\frac{0.004}{0.14} \times \$8000 = \$230$ increase in discounted lifetime earnings. Making use of the fact that roughly 40% of the students in my sample are from low-income households, the average net loss in discounted lifetime earnings is $(\$229 \times 0.4) - (\$343 \times 0.6) \approx \$115$. Pairing this with the computer cost of \$600 (maximum) per student, there is a net economic loss of \$715 per student. The results are more favorable using ELA scores: there is only an average net economic loss of \$130 per student. There could be other important costs and benefits not included in this calculation. Because teachers receive computers prior to students, there could be costs associated with training; there could be several benefits that don't show up in test scores, like improved transferable skills and better preparation for careers. In the next section, I investigate one potential benefit that is unaccounted for in this calculation: mitigation of pandemic-induced learning losses.

[19] This is an incredibly low cost compared to the \$12,000 in 2018 per-pupil spending across Wisconsin public school students that the US Census Bureau estimated.

[20] This figure is discounted at a 3% real rate back to age 12 (which is the mean age in their main analysis sample) and then inflated to 2010 dollars.

[21] This calculation first takes an average over estimated effects on elementary and middle school students, then takes the optimistic end of the 95% confidence interval.

Post-Pandemic Analysis

A multitude of recent papers have studied the impact of the COVID-19 pandemic and school closures on student achievement, finding overwhelming evidence of learning losses across the board (e.g., Aucejo et al., 2020; Engzell et al., 2021). Part of the issue was that many schools closed for several months and transitioned to virtual learning without having already provided computers to students. This made virtual learning nearly impossible initially, and forced many ill-prepared school districts to provide computers to their students at a moment's notice. In this section I ask the following question: did one-to-one initiatives implemented prior to the pandemic mitigate pandemic-related learning losses? Recall from Section 3.3 that $D = 1$ represents the group of students that received computers as part of a one-to-one initiative prior to the pandemic. I use the following econometric model to identify heterogeneous impacts of the pandemic based on when a student received a computer (prior to the pandemic or during the pandemic):

$$Y_{it} = \beta_0^P + \beta_1^P[\mathbb{1}[D = 1] \times \mathbb{1}[t \geqslant 2020]] + X'_{it}\,\beta_X^P + \gamma_{gj} + \omega_t + \varepsilon_{it} \tag{3.8}$$

On the left-hand side are students' standardized test scores (ELA and Math). On the right-hand side, I interact a group indicator with an indicator for the pandemic; β_1^P can then be interpreted as the impact of the pandemic on test scores of students that previously received computers relative to students that didn't receive a computer until the pandemic began. The covariate vector X_{it} controls for lagged test scores, subsidized lunch status, gender, race/ethnicity, attendance rate, indicators for special education enrollment and limited English proficiency, and district-average spending per pupil. Here, I include a school-grade fixed effect γ_{gj} and an academic year fixed effect ω_t and estimate the model separately for students from low-income and high-income households. Again, I cluster the standard errors at the student level and estimate the model separately for students in elementary schools and students in middle schools.

Table 3.6 presents descriptive evidence of heterogeneous impacts of the pandemic based on whether or not a student received a computer prior to the start of the pandemic. For the most part, the results are statistically insignificant. The result that does stand out is that middle school students from low-income households that received computers as part of a pre-pandemic one-to-one technology initiative fared better than their counterparts receiving computers that were part of technology initiatives forced by school closures. Though this is only a descriptive (non-causal) exercise, it does suggest that there were long-term benefits of one-to-one initiatives that could never have been predicted, such as a mitigation of pandemic-induced learning losses. Students in school-grades that had one-to-one initiatives in place prior to the pandemic likely had less of a timing gap between school closures and the start of virtual learning; in addition, they had already endured the computer-adjustment period alluded to in this paper.

Table 3.6: Heterogeneous Impacts of the Pandemic Based on
Whether a One-to-One Initiative was Already in Place (Equation (3.8))

	Outcome: ELA Test Scores		Outcome: Math Test Scores	
	Elementary School (1)	Middle School (2)	Elementary School (3)	Middle School (4)
High-income	0.001	−0.011	−0.015	−0.017
	(0.023)	(0.013)	(0.021)	(0.012)
R-squared	0.619	0.625	0.632	0.657
Student-year observations	45,807	51,623	45,823	51,677
Low-income	0.009	0.020	−0.024	0.042
	(0.031)	(0.018)	(0.038)	(0.021)
R-squared	0.589	0.617	0.557	0.510
Student-year observations	38,803	35,964	38,929	36,102
School-grade fixed effects	Y	Y	Y	Y
Academic year fixed effects	Y	Y	Y	Y

Notes: This table presents results from the two-way fixed effect regression specified in Equation (3.8) with ELA and Math test scores on the left-hand side. The model is estimated separately for students from high-income households and low-income households. All regressions include school-grade fixed effects and academic year fixed effects. In addition, all regressions include lagged test scores, an indicator for subsidized lunch eligibility, student gender, race/ethnicity, special education indicator, limited English proficiency indicator, attendance rate, and average district spending per pupil. Standard errors are clustered at the student level as with previous regressions.

3.6 Conclusion

This paper studies the effects of one-to-one technology initiatives - policies in which schools or school districts provide computers to every student - on test scores and achievement gaps using data from some of Wisconsin's largest public school districts. With one-to-one initiatives starting in 2014 - prior to the availability of standardized test scores among high school students - I focus on students in elementary and middle schools. Paying special attention to differences across students from low-income and high-income households, I start with a two-way fixed effects design and then move to an event study to analyze dynamic treatment effects.

I find that one-to-one initiatives widen ELA achievement gaps by 0.03-0.04 SD, both in elementary schools and in middle schools, an effect that is driven by statistically significant negative effects on students from low-income households. These effects are likely due to a combination of two factors: (1) technological dependence of ELA curricula as a result of one-to-one initiatives, and (2) limitations of internet and other resources at home for students from low-income households.

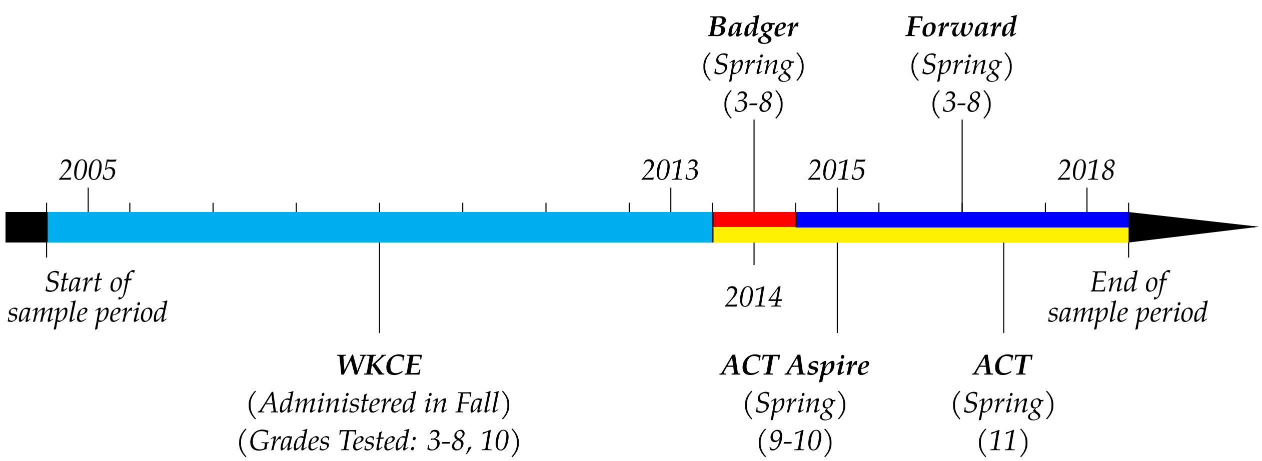
There is a slight negative impact of one-to-one initiatives on Math test scores of students from low-income and high-income households, resulting in a null effect on Math achievement gaps. These negative effects on Math test scores may be due to the computers serving as a distraction rather than a learning aid.

A simple cost-benefit analysis suggests modest economic losses of one-to-one technology initiatives, though it omits a multitude of potential benefits of the policy, including improved transferable skills, better preparation for careers, and mitigation of subsequent pandemic-induced learning losses. Altogether, there are important implications of this paper, particularly for the implementation of policies that are likely to favor some students over others. One-to-one initiatives stand a better chance of benefiting students from lower-income households and closing the "digital divide" if computers are provided with mobile hotspots; this way, students from low-income households would be able to better utilize the computers for assignments outside of school.

There is much potential for future research looking at one-to-one initiatives as well as in the broader area of technological integration with academic curricula. This is especially true, given the impact of COVID-19 on school closures, virtual learning, and losses in student achievement. Full dependence of schools on technology for an extended period followed by a return to in-person instruction has created a world in which many students are accustomed to a hybrid model of virtual and in-person learning, but many parents and teachers would prefer more separation between technology and school. More research in this area will shed light on the level of technological integration that is best for the students.

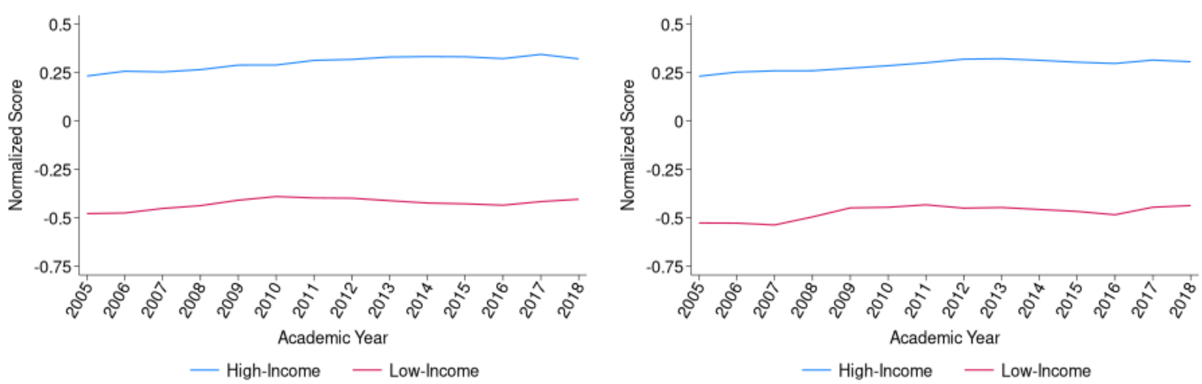
3.A Supplementary Tables and Figures

Figure 3A.1: Evolution of the Wisconsin Student Assessment System



Notes: In 2014, the WSAS expanded to include all grade levels 3-11. While Math and English Language Arts (ELA) are tested in each grade, Science and Social Studies are administered less frequently.

Figure 3A.2: Math Achievement Gaps Between Students from Low-Income Households and Students from High-Income Households



(a) Elementary school

(b) Middle school

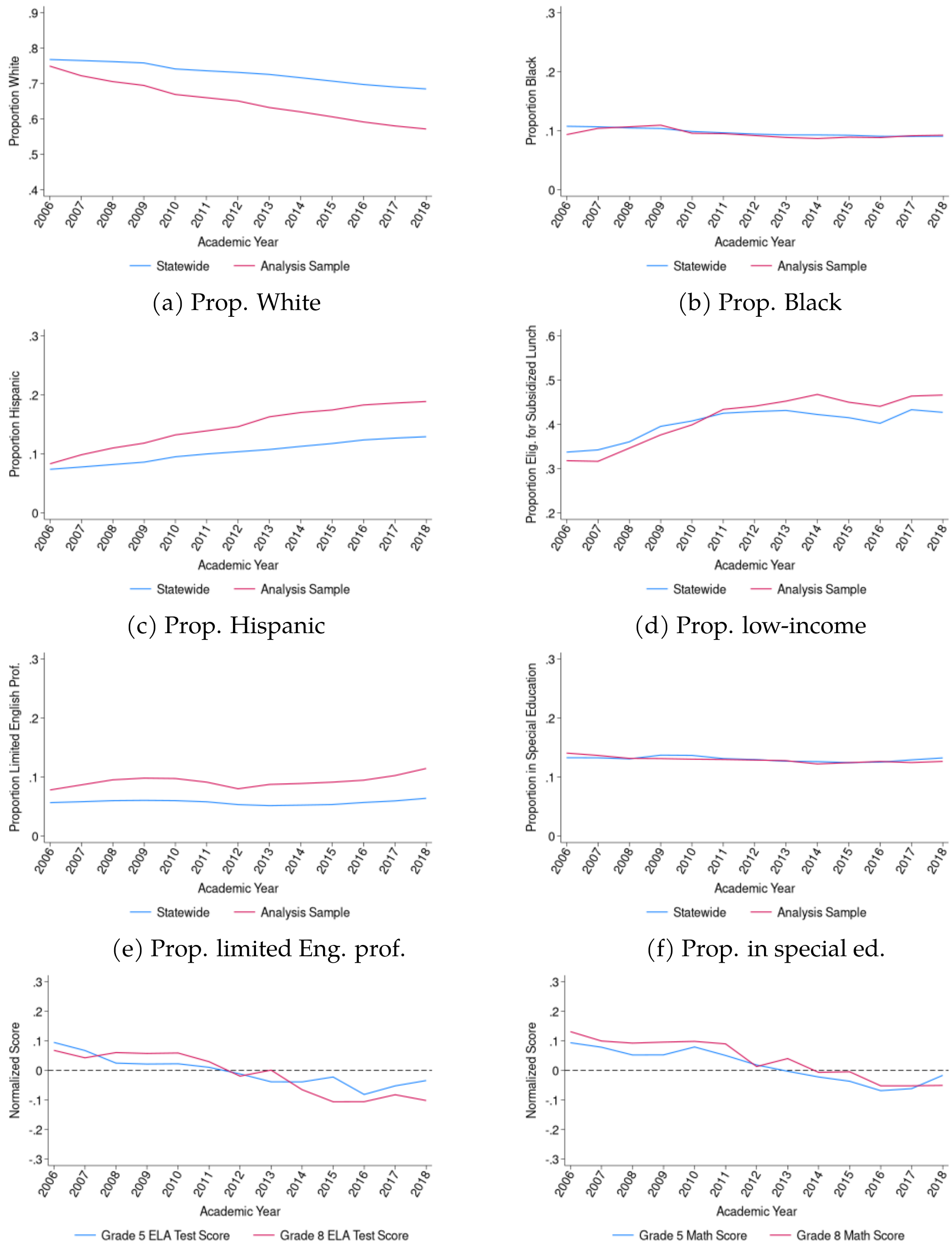
Notes: These graphs show longstanding Math achievement gaps between students from low-income and high-income households, which come from subsidized lunch eligibility. The left graph is for students from elementary schools while the right graph is for students from middle schools.

Table 3A.1: Descriptive Statistics of Students Grades 3-8 in the Sample and Statewide (2010 and 2014: Middle Years of the Analysis Window)

	2010 Academic Year		2014 Academic Year	
	Sample (1)	Statewide (2)	Sample (3)	Statewide (4)
Demographic characteristics				
White	0.657	0.741	0.617	0.716
Black	0.100	0.099	0.090	0.093
Hispanic	0.134	0.095	0.164	0.113
Male	0.511	0.511	0.515	0.513
Eligible for subsidized lunch	0.439	0.408	0.473	0.422
Limited English proficient	0.109	0.060	0.091	0.052
Enrolled in special education	0.135	0.137	0.123	0.126
Test scores				
Average grade 5 ELA score	0.029	0	−0.042	0
Average grade 5 Math score	0.104	0	−0.029	0
Average grade 8 ELA score	0.062	0	−0.104	0
Average grade 8 Math score	0.103	0	−0.051	0
Number of unique schools				
Elementary	144	872	144	838
Middle	43	739	43	805
Number of unique students	42,634	361,018	41,320	362,047

Notes: This table presents descriptive statistics for students in the analysis sample (Columns (1) and (3)) and for students in grades 3-8 statewide (Columns (2) and (4)). The restriction of grades 3-8 comes from (1) the fact that students begin taking standardized exams in grade 3, and (2) I focus on elementary and middle school students in this paper. Columns (1) and (2) are for the 2010 academic year while Columns (3) and (4) are for 2014; these are years in the middle of the analysis window. Patterns are similar here as in Table 3.2: the analysis sample is more diverse than statewide averages. In addition, the demographic composition and test scores of both the sample and the state in 2010 and 2014 are between their levels in 2006 and 2018, indicating that - while the sample is fairly representative of the state, and the composition is fairly stable over time - there is a slight trend.

Figure 3A.3: Trends in Demographic Composition of Students in the Sample and Statewide



(a) Prop. White

(b) Prop. Black

(c) Prop. Hispanic

(d) Prop. low-income

(e) Prop. limited Eng. prof.

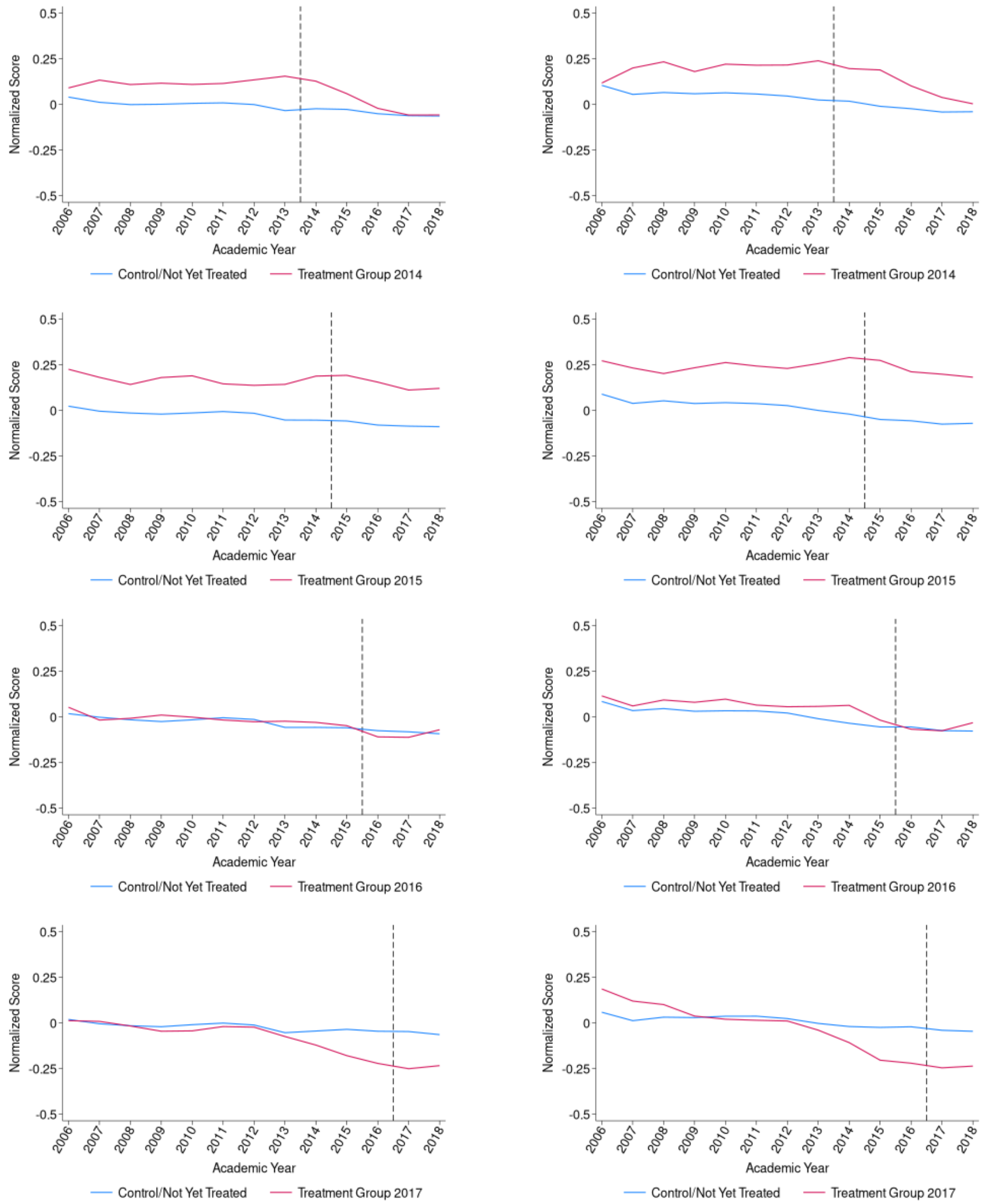
(f) Prop. in special ed.

(g) ELA test scores

(h) Math test scores

Notes: These graphs show trends in demographic composition and test scores among students grades 3-8 in the analysis sample (red line) and statewide (blue line). In panels (g) and (h), the dashed horizontal line represents the statewide mean of zero. These graphs are a supplement to the cross-sections that are presented in Tables 3.2 and 3A.1, providing a more comprehensive picture with the panel structure of the data.

Figure 3A.4: Test Score Trends (Treatment Group Divided by Year of Implementation)



(a) ELA test scores (b) Math test scores

Notes: These graphs show trends in test scores among students that received a laptop in the given year (red line) and students that either had not yet received a laptop or were not scheduled to receive one (blue line). The dashed vertical line in each graph indicates the period immediately preceding treatment (2014 in the top panel, 2015 in the second panel, etc.). Due to small samples of treated students, there is more time variation in test scores among the treated group, leading to trends that are not parallel. However, when analyzing the treatment group as a whole, trends are more stable. Elementary and middle schools are pooled here for simplicity, though they are analyzed separately.

3.B Robustness of the Two-Way Fixed Effects Design

Table 3B.1: Robustness of Equation (3.3) to the Fixed Effects Included

	Elementary School			Middle School		
	(1)	(2)	(3)	(4)	(5)	(6)
	Outcome: ELA test scores					
High-income	0.005	0.015	0.024	0.008	0.002	0.055
	(0.008)	(0.007)	(0.019)	(0.005)	(0.005)	(0.009)
Low-income	−0.032	−0.051	−0.014	−0.014	−0.030	0.043
	(0.010)	(0.009)	(0.021)	(0.006)	(0.005)	(0.009)
Achievement gap	0.037	0.066	0.037	0.022	0.032	0.012
	(0.011)	(0.011)	(0.012)	(0.006)	(0.006)	(0.007)
R-squared	0.672	0.667	0.677	0.696	0.693	0.698
Student-year observations	227,290	227,320	227,320	326,771	326,772	326,772
	Outcome: Math test scores					
High-income	−0.011	0.020	−0.182	−0.024	−0.036	−0.055
	(0.007)	(0.007)	(0.020)	(0.005)	(0.004)	(0.010)
Low-income	−0.010	−0.014	−0.208	−0.017	−0.026	−0.043
	(0.011)	(0.011)	(0.023)	(0.007)	(0.006)	(0.011)
Achievement gap	−0.001	0.034	0.026	−0.007	−0.010	−0.012
	(0.012)	(0.012)	(0.013)	(0.007)	(0.007)	(0.007)
R-squared	0.658	0.650	0.667	0.711	0.707	0.713
Student-year observations	229,068	229,099	229,099	328,081	328,081	328,081
School-grade-group fixed effects	Y			Y		
Cohort fixed effects		Y			Y	
School-cohort fixed effects			Y			Y
Academic year fixed effects	Y	Y	Y	Y	Y	Y

Notes: This table presents results from the two-way fixed effect regression specified in Equation (3.3), using cohort fixed effects (in Columns (2) and (5)) and school-cohort fixed effects (in Columns (3) and (6)) instead of the school-grade fixed effects from the main specification (in Columns (1) and (4)). All regressions include academic year fixed effects. In addition, all regressions include lagged test scores, an indicator for subsidized lunch eligibility, student gender, race/ethnicity, special education indicator, limited English proficiency indicator, attendance rate, and average district spending per pupil. The effect on achievement gaps is calculated by subtracting the estimated effect on students from low-income households from the estimated effect on students from high-income households. Standard errors are clustered at the student level. The sign of the estimated effect of one-to-one initiatives on subgroups as well as the impact on achievement gaps is stable for the most part across specifications, especially for estimates that are statistically significant in the baseline specification.

Table 3B.2: Robustness of Equation (3.3) to the Covariates Included

	Elementary School			Middle School		
	(1)	(2)	(3)	(4)	(5)	(6)
	Outcome: ELA test scores					
High-income	0.005	0.002	−0.004	0.008	0.015	0.021
	(0.008)	(0.009)	(0.010)	(0.005)	(0.006)	(0.006)
Low-income	−0.032	−0.034	−0.028	−0.014	−0.009	0.002
	(0.010)	(0.011)	(0.011)	(0.006)	(0.006)	(0.006)
Achievement gap	0.037	0.036	0.024	0.022	0.024	0.019
	(0.011)	(0.011)	(0.011)	(0.006)	(0.006)	(0.006)
R-squared	0.672	0.672	0.681	0.696	0.696	0.699
Student-year observations	227,290	227,290	227,290	326,771	326,771	326,771
	Outcome: Math test scores					
High-income	−0.011	0.012	0.002	−0.024	−0.037	−0.034
	(0.007)	(0.010)	(0.010)	(0.005)	(0.006)	(0.006)
Low-income	−0.010	0.011	−0.004	−0.017	−0.026	−0.037
	(0.011)	(0.013)	(0.013)	(0.007)	(0.007)	(0.007)
Achievement gap	−0.001	0.001	0.006	−0.007	−0.011	0.003
	(0.012)	(0.012)	(0.012)	(0.007)	(0.007)	(0.007)
R-squared	0.658	0.658	0.659	0.711	0.711	0.716
Student-year observations	229.068	229,068	229,068	328,081	328,081	328,081
School-grade-group fixed effects	Y	Y	Y	Y	Y	Y
Academic year fixed effects	Y	Y	Y	Y	Y	Y
Lagged treatment variable		Y	Y		Y	Y
Higher-order terms			Y			Y

Notes: This table presents results from the two-way fixed effect regression specified in Equation (3.3), with additional specifications for robustness. The baseline results are in Columns (1) and (4); in Columns (2) and (5), I add a lag on the treatment variable; in Columns (3) and (6), I keep the lagged treatment variable and add higher-order terms, including squared lagged test scores, squared attendance rate, and squared spending per pupil. All regressions include school-grade and academic year fixed effects. In addition, all regressions include lagged test scores, an indicator for subsidized lunch eligibility, student gender, race/ethnicity, special education indicator, limited English proficiency indicator, attendance rate, and average district spending per pupil. The effect on achievement gaps is calculated by subtracting the estimated effect on students from low-income households from the estimated effect on students from high-income households. Standard errors are clustered at the student level. For some of the estimated effects that are statistically insignificant, the sign changes across specifications, but these differences across specifications are generally statistically insignificant. Nearly all of the estimated effects that are statistically different from zero in the base specification are stable across specifications.

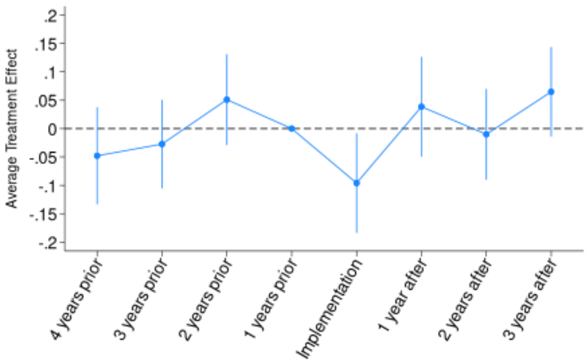
Table 3B.3: Heterogeneous Treatment Effects of One-to-One Initiatives by Grade Level

	Elementary School		Middle School		
	Grade 4 (1)	Grade 5 (2)	Grade 6 (3)	Grade 7 (4)	Grade 8 (5)
	Outcome: ELA test scores				
High-income	0.005	0.010	0.026	−0.027	0.025
	(0.011)	(0.011)	(0.009)	(0.009)	(0.010)
Low-income	−0.051	−0.008	0.004	−0.053	−0.002
	(0.015)	(0.014)	(0.010)	(0.010)	(0.011)
Achievement gap	0.056	0.018	0.022	0.026	0.027
	(0.017)	(0.016)	(0.011)	(0.003)	(0.004)
R-squared	0.654	0.690	0.695	0.694	0.694
Student-year observations	110,402	110,899	109,264	109,464	105,605
	Outcome: Math test scores				
High-income	0.106	−0.128	−0.014	0.001	−0.083
	(0.011)	(0.011)	(0.009)	(0.008)	(0.010)
Low-income	0.051	−0.067	−0.009	0.024	−0.089
	(0.016)	(0.018)	(0.011)	(0.012)	(0.014)
Achievement gap	0.056	−0.060	−0.005	−0.023	0.006
	(0.018)	(0.019)	(0.012)	(0.003)	(0.004)
R-squared	0.649	0.668	0.710	0.719	0.699
Student-year observations	111,402	111,655	109,598	109,697	106,103
School-grade-subgroup fixed effects	Y	Y	Y	Y	Y
Academic year fixed effects	Y	Y	Y	Y	Y

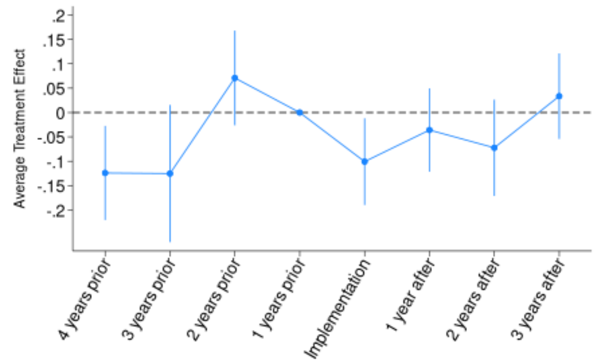
Notes: This table presents results from the two-way fixed effect regression specified in Equation (3.3) with ELA and Math test scores on the left-hand side. Instead of estimating the model separately for elementary schools and middle schools as in the main specification, here I separate further by grade level. All regressions include school-grade fixed effects and academic year fixed effects. In addition, all regressions include lagged test scores, an indicator for subsidized lunch eligibility, student gender, race/ethnicity, special education indicator, limited English proficiency indicator, attendance rate, and average district spending per pupil. The effect on achievement gaps is calculated by subtracting the estimated effect on students from low-income households from the estimated effect on students from high-income households. Standard errors are clustered at the student level. Grade 3 is omitted from the regression because I am unable to calculate lagged test scores for these students (standardized exams begin in grade 3). In elementary schools, much of the effect of one-to-one initiatives on achievement gaps comes from effects on younger students; in middle schools, the effect is fairly stable across grade levels.

3.C Robustness of the Event Study Design

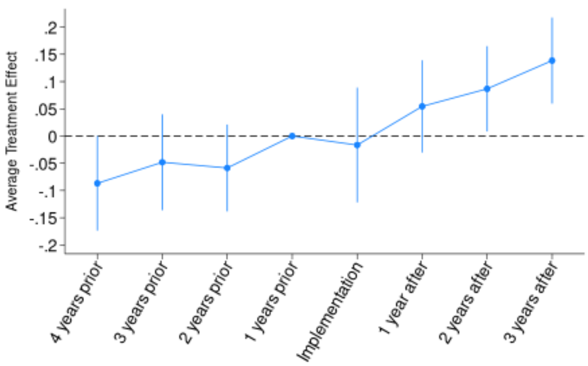
Figure 3C.1: Dynamic Effects on Test Scores of Elementary School Students
that were Given Laptops in 2014 or 2015



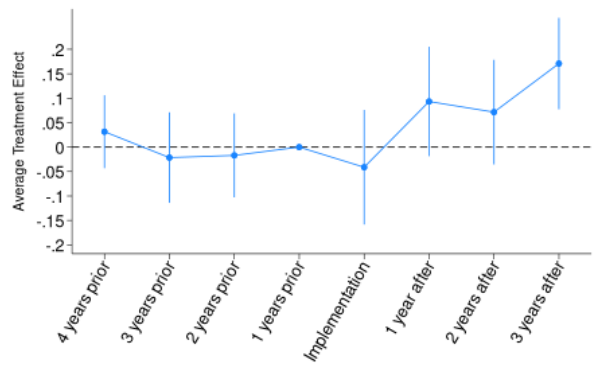
(a) Elem. ELA test scores, high-income



(b) Elem. ELA test scores, low-income



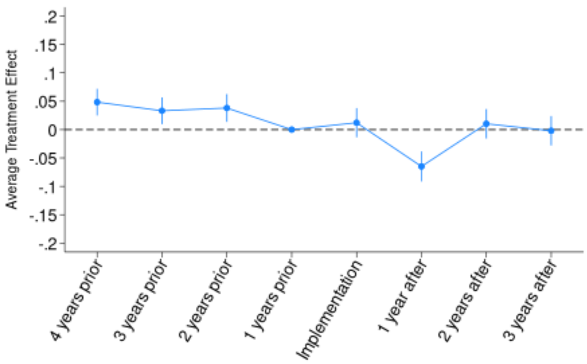
(c) Elem. Math test scores, high-income



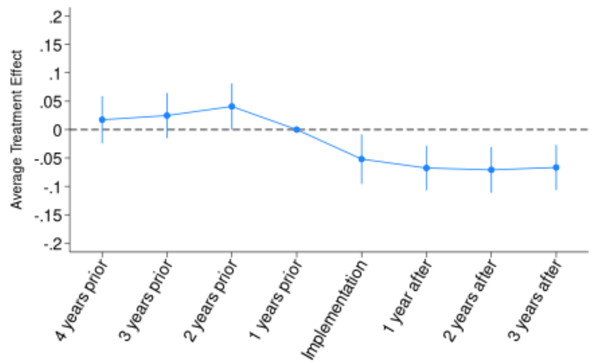
(d) Elem. Math test scores, low-income

Notes: These graphs show estimated dynamic effects on test scores of elementary school students in school-grades that received laptops in 2014 or 2015. Because the analysis window ends in 2018, these are the only treated students for whom I observe test scores three years after implementation. The worry with the main specification - and the reason for this robustness - is that the sample of treated students changes in years following implementation; findings are similar here but with more noise due to the smaller sample. Vertical lines represent 95% confidence intervals around point estimates.

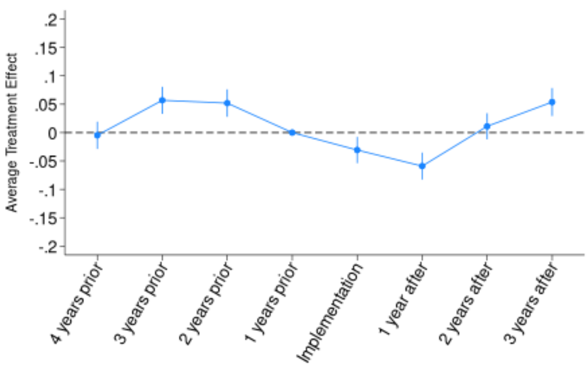
Figure 3C.2: Dynamic Effects on Test Scores of Middle School Students that were Given Laptops in 2014 or 2015



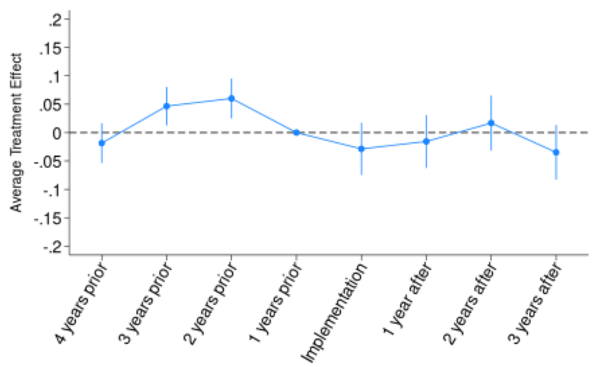
(a) Mid. ELA test scores, high-income



(b) Mid. ELA test scores, low-income



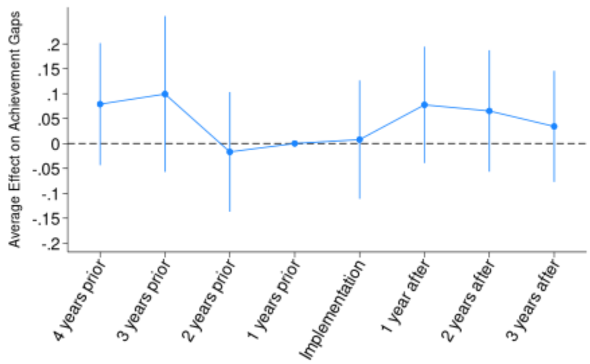
(c) Mid. Math test scores, high-income



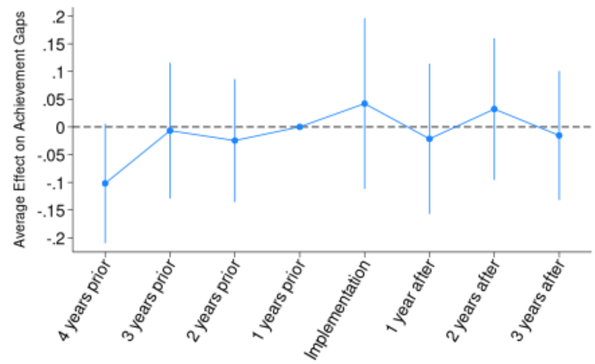
(d) Mid. Math test scores, low-income

Notes: These graphs show estimated dynamic effects on test scores of middle students in school-grades that received laptops in 2014 or 2015. Because the analysis window ends in 2018, these are the only treated students for whom I observe test scores three years after implementation. The worry with the main specification - and the reason for this robustness - is that the sample of treated students changes in years following implementation; findings are similar here but with more noise due to the smaller sample. Vertical lines represent 95% confidence intervals.

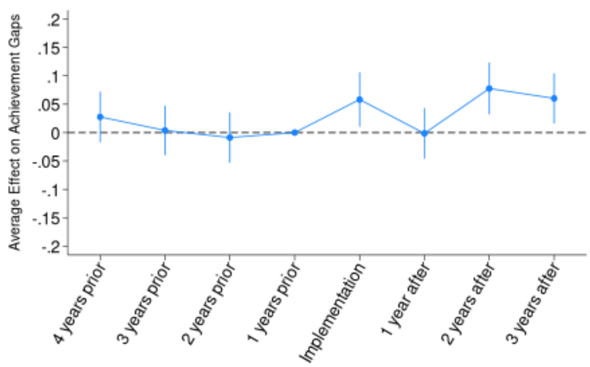
Figure 3C.3: Dynamic Effects on Income-Based Achievement Gaps
Among Students that were Given Laptops in 2014 or 2015



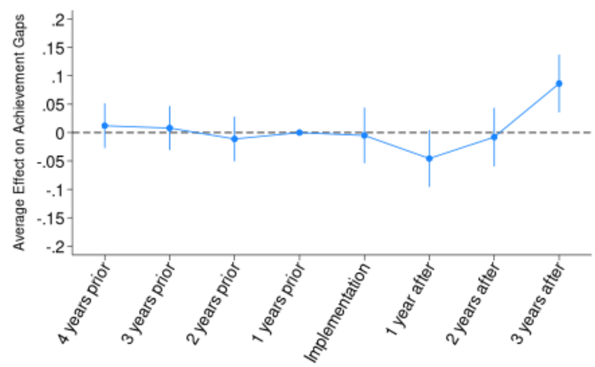
(a) ELA test scores, elementary



(b) Math test scores, elementary



(c) ELA test scores, middle



(d) Math test scores, middle

Notes: These graphs show estimated dynamic effects of one-to-one initiatives on achievement gaps between students from low-income and high-income households; these come from subsidized lunch eligibility. Here I only include students in school-grades that received laptops in 2014 or 2015. Because the analysis window ends in 2018, these are the only treated students for whom I observe test scores three years after implementation. The worry with the main specification - and the reason for this robustness - is that the sample of treated students changes in years following implementation; findings are similar. The top set of graphs are for the sample of elementary school students while the bottom set are for middle school students. The left set of graphs use ELA test scores as the outcome while the right set use Math test scores. Vertical lines represent 95% confidence intervals.

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and student achievement in the chicago public high schools. *Journal of Labor Economics* 25(1):95–135.

Abdulkadiroğlu, Atila, Parag A. Pathak, Jonathan Schellenberg, and Christopher R. Walters. 2020. Do Parents Value School Effectiveness? *American Economic Review* 110(5):1502–1539.

Abraham, Sarah, and Liyang Sun. 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics* 225(2):175–199.

Adams, E. Kathleen, Janet M. Bronstein, and Cheryl Raskind-Hood. 2002. Adjusted clinical groups: Predictive accuracy for medicaid enrollees in three states. *Health Care Financing Review* 24(1):43–61.

Ahn, Sangtae, and Jeffrey A. Fessler. 2003. Standard errors of mean, variance, and standard deviation estimators. Tech. Rep., University of Michigan, EECS Department.

Ahn, Tom, Esteban Aucejo, and Jonathan James. 2023. The importance of matching effects for labor productivity: Evidence from teacher-student interactions. Working Paper 2106, California Polytechnic State University, Department of Economics.

Aizer, Anna. 2007. Public health insurance, program take-up, and child health. *The Review of Economics and Statistics* 89(3):400–415.

Almond, Douglas, Janet Currie, and Valentina Duque. 2018. Childhood Circumstances and Adult Outcoems: Act II. *Journal of Economic Literature* 56(4):1360–1446.

Altonji, J. 1995. The effects of high school curriculum on education and labor market outcomes. *The Journal of Human Resources* 30(3):3–32.

Altonji, Joseph G., Todd Elder, and Christopher Taber. 2005. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy* 113(1):151–184.

Altonji, Joseph G., Todd Elder, and Christopher Taber. 2008. Using selection on observed variables to assess bias from unobservables when evaluating swan-ganz catheterization. *American Economic Review: Papers & Proceedings* 98(2):345–350.

Angrist, Josh, Peter Hull, and Christopher Walters. 2023. Methods for measuring school effectiveness. *Handbook of the Economics of Education* 7:1–60.

Angrist, Josh, and Victor Lavy. 2002. New evidence on classroom computers and pupil learning. *The Economic Journal* 112:735–765.

Angrist, Joshua, Peter Hull, Parag Pathak, and Christopher Walters. 2016. Interpreting tests of school vam validity. *American Economic Review* 106(5):388–92.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. 2017. Leveraging Lotteries for School Value-Added: Testing and Estimation. *Quarterly Journal of Economics* 132(2): 871–919.

Arce-Trigatti, Paula. 2018. The impact of state-mandated advanced placement programs on student outcomes. *Economics of Education Review* 63(C):180–193.

Athey, Susan, and Stefan Wager. 2019. Estimating treatment effects with causal forests: An application. *Observational Studies* 5(2):37–51.

Aucejo, Esteban, Patrick Coate, Jane C. Fruehwirth, Sean Kelly, and Zachary Mozenter. 2022. Teacher effectiveness and classroom composition: Understanding match effects in the classroom. *The Economic Journal* 132(648):3047–3064.

Aucejo, Esteban, Jacob French, Maria Paola Ugalde Araya, and Basit Zafar. 2020. The impact of covid-19 on student experiences and expectations: Evidence from a survey. *Journal of Public Economics* 191:1–15.

Aughinbaugh, Alison. 2012. The effects of high school math curriculum on college attendance: Evidence from the nlsy97. *Economics of Education Review* 31(6):861–870.

Aypay, Ahmet. 2010. Information and communication technology (ict) usage and achievement of turkish students in pisa 2006. *Turkish Online Journal of Educational Technology* 9(2):116–124.

Bacher-Hicks, Andrew, and Cory Koedel. 2023. Estimation and interpretation of teacher value added in research applications. *Handbook of the Economics of Education* 6C:93–134.

Barrera-Osorio, Felipe, and Leigh Linden. 2009. The use and misuse of computers in education: Evidence from a randomized experiment in colombia. Working Paper 4836, World Bank Policy Research.

Belo, Rodrigo, Pedro Ferreira, and Rahul Telang. 2014. Broadband in school: Impact on student performance. *Management Science* 60(2):265–282.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. How much should we trust difference-in-differences estimates? *The Quarterly Journal of Economics* 119(1):249–275.

Beuermann, Diether, Julian Cristia, Santiago Cueto, Ofer Malamud, and Yyannu Cruz-Aguayo. 2015. One laptop per child at home: Short-term impacts from a randomized experiment in peru. *American Economic Journal: Applied Economics* 7(2):53–80.

Bharadwaj, Prashant, Juan Pedro Eberhard, and Christopher A. Neilson. 2018. Health at Birth, Parental Investments, and Academic Outcomes. *Journal of Labor Economics* 36(2):349–394.

Biau, G, and E Scornet. 2016. A random forest guided tour. *Test* 25(2):197–227.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2024. Revisiting event-study designs: Robust and efficient estimation. *The Review of Economic Studies* 91(6):3253–3285.

Branch, Gregory, Eric Hanushek, and Steven Rivkin. 2012. Estimating the effect of leaders on public sector productivity: The case of school principals. Working Paper 17803, National Bureau of Economic Research.

Breiman, Leo. 2001. Random forests. *Machine Learning* 45:5–32.

de Buhr, Elke, Michael Ewers, and Antje Tannen. 2020. Potentials of school nursing for strengthening the health literacy of children, parents, and teachers. *International Journal of Environmental Research and Public Health* 17(7).

Bulman, George, and Robert Fairlie. 2016. Technology and education: Computers, software, and the internet. *Handbook of the Economics of Education* 5:239–280.

Bulut, Okan, and Erhan Delen. 2011. The relationship between students' exposure to technology and their achievement in science and math. *Turkish Online Journal of Educational Technology* 10(3): 311–317.

Case, Anne, Angela Fertig, and Christina Paxson. 2005. The lasting impact of childhood health and circumstance. *Journal of Health Economics* 24(2):365–389.

de Chaisemartin, Clément, and Xavier D'Haultfoeuille. 2020. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review* 110(9):2964–2996.

Chatterji, Roby, Neil Campbell, and Abby Quirk. 2021. Closing advanced coursework equity gaps for all students. Tech. Rep., Center for American Progress.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review* 104(9):2593–2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review* 104(9): 2633–2679.

Cristia, Julian, Alejo Czerwonko, and Pablo Garofalo. 2014. Does technology in schools affect repetition, dropout and enrollment? evidence from peru. *Journal of Applied Economics* 17(1):89–111.

Cristia, Julian, Pablo Ibarrarán, Santiago Cueto, Ana Santiago, and Euginio Severín. 2017. Technology and child development: Evidence from the one laptop per child program. *American Economic Journal: Applied Economics* 9(3):295–320.

Currie, Janet. 2009. Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development. *Journal of Economic Literature* 47(1):87–122.

Currie, Janet, Mark Stabile, Phongsack Manivong, and Leslie L. Roos. 2010. Child Health and Young Adult Outcomes. *Journal of Human Resources* 45(3):517–548.

Currie, Janet, and Duncan Thomas. 2000. School quality and the longer-term effects of head start. *Journal of Human Resources* 35(4):755–774.

Dague, Laura, and Rebecca Myerson. 2024. Loss of medicaid coverage during the renewal process. *Journal of the American Medical Association Health Forum* 5(5).

Deming, David. 2009. Early Childhood Interventioan and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics* 1(3):111–134.

Deming, David. 2014. Using school choice lotteries to test measures of school effectiveness. *American Economic Review: Papers & Proceedings* 104:406–411.

Deming, David, Justine Hastings, and Thomas Kane Douglas Staiger. 2014. School choice, school quality, and postsecondary attainment. *American Economic Review* 104(3):991–1013.

Dynarski, Susan M., and Katherine Michelmore. 2016. The gap within the gap: Finding a better measure of student disadvantage. Tech. Rep., Education Policy Initiative, University of Michigan.

Einav, Liran, Amy Finkelstein, and Neale Mahoney. 2022. Producing health: Measuring value added of nursing homes. Working Paper 30228, National Bureau of Economic Research.

Engzell, Per, Arun Frey, and Mark Verhagen. 2021. Learning loss due to school closures during the covid-19 pandemic. *Proceedings of the National Academy of Sciences of the United States of America* 118(17).

Escueta, Maya, Andre Joshua Nickow, Philip Oreopoulos, and Vincent Quan. 2020. Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature* 58(4):897–996.

Fairlie, Robert, and Jonathan Robinson. 2013. Experimental evidence on the effects of home computers on academic achievement among schoolchildren. *American Economic Journal: Applied Economics* 5(3):211–240.

Figlio, David, Jonathan Guryan, Krzysztof Karbownik, and Jeffrey Roth. 2014. The Effects of Poor Neonatal Health on Children's Cognitive Development. *American Economic Review* 104(12): 3921–3955.

Figlio, David, and Marianne Page. 2002. School choice and the distributional effects of ability tracking: Does separation increase inequality? *Journal of Urban Economics* 51(3):497–514.

Gershenson, Seth, Cassandra M. D. Hart, Joshua Hyman, Constance A. Lindsay, and Nicholas W. Papageorge. 2022. The long-run impacts of same-race teachers. *American Economic Journal: Economic Policy* 14(4):300–342.

Gilraine, Michael, Uros Petronijevic, and John Singleton. 2023. School choice, competition, and aggregate school quality. Working Paper 31328, National Bureau of Economic Research.

Goodman-Bacon, Andrew. 2021. Difference-in-differences with variation in treatment timing. *Journal of Econometrics* 225(2):254–277.

Grimes, Douglas, and Mark Warschauer. 2008. Learning with laptops: A multi-method case study. *Journal of Educational Computing Research* 38(3):305–332.

Grossman, Michael. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy* 80(2):223–255.

Hanushek, Eric A. 2009. Teacher deselection. *Creating a New Teaching Profession* Urban Institute Press:165–180.

Hanushek, Eric A., Paul E. Peterson, Laura M. Talpey, and Ludger Wößmann. 2022. Long-run trends in the u.s. ses-achievement gap. *Education Finance and Policy* 17(4):608–640.

Hanushek, Eric A., and Ludger Wößmann. 2006. Does educational tracking affect performance and inequality? difference-in-differences evidence across countries. *The Economic Journal* 116(510): C63–C76.

Hastie, T., R. Tibshirani, and J.H. Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction (2nd edition)*. Springer.

Hausman, Jerry. 2001. Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic Perspectives* 15(4):57–67.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. Characterizing selection bias using experimental data. *Econometrica* 66(5):1017–1098.

Heckman, James, and Xuesong Li. 2004. Selection bias, comparative advantage and heterogeneous returns to education: evidence from china in 2000. *Pacific Economic Review* 9(3):155–171.

Hockman, William. 1970. Advanced placement: Can it change with the times? *College Board Review* 77:16–19.

Jackson, C. Kirabo. 2010. A little now for a lot later: A look at a texas advanced placement incentive program. *The Journal of Human Resources* 45(3):591–639.

Jackson, Kirabo, and Alexey Makarin. 2018. Can online off-the-shelf lessons improve student outcomes? evidence from a field experiment. *American Economic Journal: Economic Policy* 10(3): 226–254.

Jia, Ning. 2021. Do stricter high school math requirements raise college stem attainment? *Economics of Education Review* 83:102–140.

Kane, Thomas, Jonah Rockoff, and Douglas Staiger. 2008. What does certification tell us about teacher effectiveness? evidence from new york city. *Economics of Education Review* 27(6):615–631.

Kane, Thomas, and Douglas Staiger. 2002. The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16(4):91–114.

Kim, Sunha, and Mido Chang. 2010. Does computer use promote the mathematical proficiency of ell students? *Journal of Educational Computing Research* 42(3):285–305.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review: Papers & Proceedings* 105(5):491–95.

Kozma, Robert. 2003. Technology and classroom practices: An international study. *Journal of Research on Technology in Education* 36(1):1–14.

Lee, Jihyun, Wendy Grigg, and Gloria Dion. 2007. The nation's report card: Mathematics 2007. Technical Report, National Center for Education Statistics, Institute of Education Sciences, US Department of Education.

Machin, Stephen, Sandra McNally, and Olmo Silva. 2007. New technology in schools: Is there a payoff? *The Economic Journal* 117(522):1145–1167.

Martin, Michael. 2021. Computer and internet use in the united states: 2018. Technical Report, United States Census Bureau.

Means, Barbara, William Penuel, and Christine Padilla. 2001. *The connected school: Technology and learning in high school*. Jossey-Bass Education Series.

Mullainathan, Sendhil, and Jann Spiess. 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives* 31(2):87–106.

Obermeyer, Ziad, and Ezekiel Emanuel. 2016. Predicting the future - big data, machine learning, and clinical medicine. *New England Journal of Medicine* 375(13):1216–1219.

Owen, Stephanie. 2024. The advanced placement program and educational inequality. *Education Finance and Policy* 1–32.

Papanastasiou, Elena, Michalinos Zembylas, and Vrasidas Charalambos. 2003. Can computer use hurt science achievement? the usa results from pisa. *Journal of Science Education and Technology* 12(3):325–332.

Patterson, Richard. 2018. Can behavioral tools improve online student outcomes? experimental evidence from a massive open online course. *Journal of Economic Behavior and Organization* 153: 293–321.

Penuel, William. 2006. Implementation and effects of one-to-one computing initiatives: A research synthesis. *Journal of Research on Technology in Education* 38(3):329–348.

Rothschild, Eric. 1999. Four decades of the advanced placement program. *The History Teacher* 32(2): 175–206.

Rothstein, Jesse M. 2006. Good principals or good peers? parental valuation of school characteristics, tiebout equilibrium, and the incentive effects of competition among jurisdictions. *American Economic Review* 96(4):1333–1350.

Rothstein, Jesse M. 2010. Teacher Quality in Educational Production: Tracking, Decay, and Student Acheivement. *Quarterly Journal of Economics* 125(1):175–214.

Rothstein, Jesse M. 2017. Measuring the Impacts of Teachers: Comment. *American Economic Review* 107(6):1656–1684.

Schneider, Jack. 2009. Privilege, equity, and the advanced placement program: Tug of war. *Journal of Curriculum Studies* 41(6):813–831.

Schofield, Janet, and Ann Davidson. 2002. *Bringing the internet to school: Lessons from an urban district*. Jossey-Bass Education Series.

Shatte, Adrian, Delyse Hutchinson, and Samantha Teague. 2019. Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine* 49(9):1426–1448.

Steinberg, Matthew P., and Rachel Garrett. 2016. Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis* 38(2):293–317.

Suhr, Kurt, David Hernandez, Douglas Grimes, and Mark Warschauer. 2010. Laptops and fourth-grade literacy: Assisting the jump over the fourth-grade slump. *The Journal of Technology, Learning, and Assessment* 9(5):1–45.

Sweet, Richard, and Alina Meates. 2004. Ict and low achievers: What does pisa tell us? Technical Report, Promoting Equity Through ICT in Education: Projects, Problems, Prospects.

Tan, Joshua, Xiaojin Zhang, Dawn Cheng, Ian Yi Onn Leong, Chia Siong Wong, Jeannie Tey, Shu Ching Loh, Eugene Fidelis Soh, and Wei Yen Lim. 2023. Using the johns hopkins acg case-mix system for population segmentation in a hospital-based adult patient population in singapore. *BMJ Open* 13(3):e062786.

Walters, Christopher. 2015. Inputs in the production of early childhood human capital: Evidence from head start. *American Economic Journal: Applied Economics* 7(4):76–102.

Weiner, J P, Allen Dobson, Stephanie Maxwell, Kevin Coleman, Barbara Starfield, and Gerard Anderson. 1996. Risk-adjusted medicare capitation rates using ambulatory and inpatient diagnoses. *Health Care Financing Review* 17(3):77–99.

Weiner, J P, B H Starfield, and R N Lieberman. 1992. Johns hopkins ambulatory care groups (acgs). a case-mix system for ur, qa, and capitation adjustment. *HMO Practice* 6(1):13–19.

Weiner, J P, B H Starfield, D M Steinwachs, and L M Mumford. 1991. Development and application of a population-oriented measure of ambulatory care case-mix. *Medical Care* 29(5):452–472.

Willis, Robert, and Sherwin Rosen. 1979. Education and self-selection. *Journal of Political Economy* 87(5):S7–S36.

Wright, S. Paul, Sandra P. Horn, and William L. Sanders. 1997. Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education* 11:57–67.