NATIONAL
CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

CALDER

AIR

Assessing the Accuracy of Elementary School Test Scores as Predictors of Students' High School Outcomes

Dan Goldhaber
Malcolm Wolff
Timothy Daly

WORKING PAPER No.235-0520-2 • August 2021

Assessing the Accuracy of Elementary School Test Scores as Predictors of Students' High School Outcomes

Dan Goldhaber
American Institutes for Research/CALDER
University of Washington

Malcolm Wolff
University of Washington

Timothy Daly
EdNavigator

Contents

Acknowledgments

This work is supported by the Carnegie Foundation and the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). All opinions expressed in this paper are those of the authors and do not necessarily reflect the views of our funders or the institutions to which the authors are affiliated. CALDER working papers have not undergone final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication. We thank Nathaniel Marcuson and Trevor Gratz for their excellent research assistance, Kathryn Choate and Trevor Gratz for suggested edits, and James Cowan and David Keeling for comments on an earlier draft. Any opinions, findings, and conclusions expressed in these papers are those of the authors and do not necessarily reflect the views of our funders.

CALDER • American Institutes for Research
1400 Crystal Drive, 10th Floor, Arlington, VA 22202
202-403-5796 • www.caldercenter.org

Assessing the Accuracy of Elementary School Test Scores as Predictors of Students' High School Outcomes

Dan Goldhaber, Malcolm Wolff, Timothy Daly

CALDER Working Paper No. 235-0520-2
August 2021

Abstract

Testing students and using test information to hold schools and, in some cases, teachers accountable for student achievement has arguably been the primary national strategy for school improvement over the past decade and a half. Tests are also used for diagnostic purposes, such as to predict students at-risk of dropping out of high school. But there is policy debate about the efficacy of this usage, in part because of disagreements about whether tests are an important schooling outcome. We use panel data from three states – North Carolina, Massachusetts and Washington State – to investigate how accurate early test scores are in predicting later high school outcomes: 10th grade test achievement, the probability of taking advanced math courses in high school, and graduation. We find 3rd grade tests predict all of these outcomes with a high degree of accuracy and relatively little diminishment from using 8th grade tests. We also find evidence that using a two-stage model estimated on separate cohorts (one predicting 8th grade information using 3rd grade information, and another predicting high school outcomes with 8th grade information) only slightly diminishes forecast accuracy. Finally, the use of machine learning techniques increases accuracy of predictions over widely used linear models, but only marginally.

1. Introduction

Testing students annually and using the results to inform policy decisions, including school accountability, has been one of the primary federal and state strategies for identifying and addressing educational inequity over the past two decades.[1] There is a considerable amount of policy disagreement about the value of tests for accountability, identifying learning gaps, and/or the amount of testing that occurs (e.g., Forte, 2021; Koretz, 2017; Strauss, 2015). Some of this is likely due to doubts about the extent to which tests at one point in time predict both later test and non-test outcomes (Goldhaber and Özek, 2019).

Tests are also intended to be useful as informative and diagnostic tools for educators and parents, allowing schools to communicate about student needs and tailor instruction and interventions. Several states and localities use state assessments in early warning systems to predict whether students are at-risk of not meeting specified academic outcomes. But existing systems rely on predictions that typically only span a few grades and do not harness the potential for long panels of data to assess high school outcomes based on elementary test achievement.[2]

Understanding the degree to which we ought to rely on test scores either for accountability or diagnostic purposes depends fundamentally on their predictive power. Predictions of long-term student outcomes are useful only if they are reasonably accurate: imprecision in predictions could result in poor allocation of school resources, for remediation, for instance. And, in terms of providing parents and families with information, inaccurate information might falsely reassure them about their children's future schooling prospects or cause unneeded concerns.[3]

In this paper, we use panel data from three states – North Carolina, Massachusetts, and Washington – to investigate how accurate early measures of student achievement are in predicting later high school outcomes. We contribute to the literature in four distinct ways. First, the long panels we employ allow us to quantify the accuracy of models predicting how early (3rd and 4th grades) measures of student background and achievement predict several later schooling outcomes including high school test achievement, high school course-taking, and high school graduation. Second, we test the extent to which predictions based on distinct segments of student data (e.g., 3rd to 8th grade, then 8th to 12th) sacrifice forecast accuracy (which is of particular policy relevance for states or localities that do not yet have long administrative data panels). Third, we test the degree to which the use of parameter estimates from models predicting schooling outcomes derived from one state diminish the accuracy of predicting outcomes in

[1] Test-based accountability predated the 2001 passage of the federal No Child Left Behind (NCLB) Act, but NCLB made it mandatory for states to use tests for a variety of accountability and reporting purposes. For more on accountability and its effects, see Figlio and Loeb (2011).

[2] Chicago (Allensworth, 2013) and Massachusetts (Curtin et al., 2012) are good examples of such systems. Limitations of early warning systems have been driven by the underlying data available for the predictions, which often has not included test scores that span early elementary years through high school. But this is rapidly changing. All states since 2005-06 have been testing students annually in math and reading in 3rd through 8th grade as a result of the No Child Left Behind Act passed in 2002 (Le Floch et al., 2007).

[3] A related issue, which arises irrespective of the accuracy of the information, is whether the provision of information itself might adversely affect student achievement. Dee (2014), for instance, finds evidence from a framed field experiment that primed awareness of negative student-athlete stereotypes reduced athlete test scores by 12% relative to non-athletes.

other states. Finally, we compare the accuracy of long-term student outcome predictions using machine learning techniques to generalized linear models.

Using generalized linear and nonlinear models, we find that students' 3rd grade test scores predict their high school outcomes nearly as accurately as their 8th grade test scores. For instance, educational achievement models based on 8th grade test scores and demographics correctly classify 70% of high school graduates while misclassifying only 28% of non-graduates. That is, the models identify most of the struggling students who will fail to graduate high school without incorrectly identifying very many who will eventually graduate. When 3rd grade test scores are used in place of middle school tests, there is little degradation of the accuracy of the predictive models, correctly classifying 65% of graduates with the same misclassification rate, suggesting that the trajectory of student achievement tends to change little from 3rd to 8th grade. While Allensworth and Easton (2007) find that 9th grade characteristics correctly classify 85% of graduates while misclassifying 28% of non-graduates, a larger panel of student information, such as GPA, credit completion and number of course failures, is used to do so. Machine learning improves average prediction accuracy, but only slightly, from one to 10% according to under the curve (AUC) measures depending on the method, outcome in question, and the elementary grades included. And the student outcome predictions based on machine learning techniques are highly correlated (over .85) with those generated by more widely used generalized linear models.

We also find that predictive models travel across state lines. That is, we can use student achievement data and parameters from one state as the basis for predicting students' educational outcomes in another state without substantially degrading forecast accuracy. As an example, in terms of predicting the likelihood of high school graduation in Massachusetts, our findings show that 3rd grade test scores accurately classify 71% of graduates while misclassifying 28% of non-graduates using prediction parameters estimated from Massachusetts data. If instead we use prediction parameters that are based on data on Washington students, the accuracy is only slightly degraded: 70% of Massachusetts graduates are accurately classified and 28% non-graduates are misclassified.[4] This analysis suggests that after knowing a student's personal characteristics and 3rd grade achievement levels, relatively little may be gained by knowing the state in which the student attends school.

Finally, consistent with existing evidence (e.g., Austin et al., 2020; Lee, 2002; Reardon, 2016), poverty and race/ethnicity are strongly predictive of high school outcomes controlling for students' elementary test achievement, and the magnitude of these demographic variables are educationally meaningful. Our models provide yet more bracing evidence of the extraordinary challenges faced by students of different backgrounds even when they display the same levels of academic mastery. For instance, an economically disadvantaged student (EDS)[5] in 3rd grade lowers the student's predicted position in the high school math distribution by 5.8 percentile points, the predicted probability of taking an advanced course in high school by 9.7 percentile

[4] In other words, with a misclassification rate of 28%, cross-state estimates correctly classify as little as two percent fewer graduates compared to in-state estimates.

[5] In response to a directive from the North Carolina Education Research and Data Center, we use the term economically disadvantaged student (EDS) to refer to students who qualify for free- or reduced-price meals. Thus, we use the term EDS throughout for all states.

points, and the predicted probability of graduation by 10.2 percentage points. Put another way, an EDS in the 3rd grade who is scoring in the highest decile in 3rd grade math test distribution has roughly an equal chance of graduating as a non EDS scoring in the second lowest decile. There are similar gaps in graduation probabilities between white and underrepresented minority students. It is not only less common for low income and minority students to reach high levels of achievement – it is more difficult to sustain those levels.

2. Literature on Predicting Long-Term Student Outcomes

A number of studies have looked at the degree to which early cognitive and non-cognitive student characteristics, including measures of student achievement and engagement predict long-term outcomes such as later test achievement, high school course-taking and graduation, and college-going and labor market earnings.[6] But few studies rely on statewide administrative data that span elementary grades through high school. The primary reason is that, until recently, only a few states had the data infrastructure necessary to reliably link students longitudinally over a long grade span.[7] And today more than half of the states still do not have easy access to detailed longitudinal data spanning 3rd grade to graduation (Data Quality Campaign, 2016).

Yet a small body of research highlights the value of such data collection for predicting long-term student outcomes. Hernandez (2011), for instance, reports summary statistics from a longitudinal study of nearly 4,000 students and finds that those who don't read proficiently by 3rd grade are four times more likely not to graduate high school on time, with the risk highest for the lowest performers, and the effect even more pronounced for EDS status students.[8] This widely-cited report was influential in shaping federal and state early reading intervention strategies; at least four states have passed 3rd grade reading laws since the report's release,[9] while other states have amended their 3rd grade reading laws multiple times or phased in various requirements (CCSSO, 2010).

Other studies rely on shorter panels but illustrate the importance of early academic indicators in predicting future academic success. Goldhaber et al. (2018), for instance, finds significant evidence that 3rd grade test scores are strongly predictive of 8th grade test outcomes as well as high school math and science course-taking patterns. Two other recent studies show that students' high school GPA is a strong predictor of high school graduation as well as college-

[6] See Murnane et al. (1995), and Cawley et al. (2001), Cunha & Heckman (2006), Heckman et al. (2006), Todd & Wolpin (2007), and Cunha et al. (2010).

[7] The Data Quality Campaign, which has been tracking the extent to which states collect "10 Essential Elements of Statewide Longitudinal Data Systems" considered necessary to build a highly effective longitudinal data system (Data Quality Campaign, 2009). In 2005, fewer than 8 states recorded all elements and less than half of states had an audit system in place to assess data reliability. Most lacked information on courses completed, grades earned, and student-level college readiness test scores such as Advanced Placement (AP) tests. By 2011 most states met the 10 essential elements, where at most 9 states did not meet the requirement.

[8] 23% of the lowest performing readers do not graduate high school on time, relative to nine percent for *basic* readers and four percent for *proficient* readers. Furthermore, not only are children who have lived in poverty 3.7 times more likely to not graduate from high school, but the lowest performing readers in this group are 6 times more likely than proficient readers to fail to graduate high school on time.

[9] Several more states brought bills into consideration that ultimately did not pass.

going, retention, and graduation (Geiser & Stanelices, 2007; Easton et al., 2017).[10] Similarly, Silver et al. (2008) follow a cohort of Los Angeles Unified School District students over a seven year period and find that test scores as early as 6th grade are predictive of on-time high school graduation.

Zau and Betts (2008) address the feasibility of using earlier elementary indicators for predicting long-term academic achievement. Using administrative data to predict the likelihood that students pass the California High School Exit Exam (CAHSEE), a formerly required component of California's school accountability program,[11] the authors find evidence suggesting predictions using 4th grade test scores and student characteristics have nearly the same accuracy as predictions using the same metrics observed in 9th grade, highlighting administrators' ability to easily identify and provide assistance to at-risk students as early as elementary school. Furthermore, the authors' results "strongly suggest eleventh-hour interventions by themselves are unlikely to yield intended results,"[12] raising a general concern for the time necessary for successful intervention. While promising, this study is primarily limited by its scope of outcome data; only one test-based outcome is considered, and the CAHSEE is written for a comprehension level of 10th grade English and 8th grade math—having closer comparability to elementary and middle school tests than other high school standardized testing, such as the SAT.

Several studies have also illustrated the importance of the scope of student input data, correlating broader indicators of school attachment or academic success with students' long-run outcomes. For example, a 2007 study by Neild et al. showed that 75% of Philadelphia 6th graders with either a final grade of "F" in math or English, below an 80% annual attendance rate, or an unsatisfactory behavior mark eventually dropped out. In addition, Allensworth and Easton (2005) find that students having accumulated five full course credits with no more than one semester "F" in a core subject is more indicative of high school graduation than standardized test scores for Chicago Public Schools students.[13]

The growth of both data availability and interest among parents, teachers, and policymakers have inspired proprietary work to begin developing methods for providing schools with accurate early warning indicators. Sorenson (2018) uses Support Vector Machines, Boosted Regression and Post-LASSO to explore the classification accuracy of eventual dropout for 9th graders participating in the High School Longitudinal Study survey of 2009, accurately classifying 89-90% of dropouts while misclassifying 16-28% of dropouts. With over six million student-year observations from 6th to 12th grade across 32 states, Christie et al. (2019) use gradient-boosted decision trees to predict the risk of dropout of students. The authors use a wide range of current and historical yearly predictors, including attendance, academic performance, behavior, household and enrollment stability, and other contextual information, finding high dropout prediction accuracy. However, the primary pitfall of such machine learning algorithms is an inability to assess the contribution of each predictor independently—an important aspect of

[10] They also find that overall high school GPAs are highly correlated between freshman and junior year, suggesting the ability to predict future outcomes in as early as 9th grade.

[11] The CAHSEE was suspended effective January 1, 2016.

[12] They find of those in San Diego who failed to graduate in spring 2006 because of the CAHSEE, only 27% re-enrolled the next year, and only 3.1 percent passed in the following year.

[13] Core subjects include English, math, science, or social studies.

communicating results and actionable solutions to families. The methodology may accurately classify a young student as, for instance, a dropout risk, but this may be of limited value as it does not provide much information about precisely why students are at-risk of dropping out. Similarly, Sorenson (2019) explores decision tree methodologies to assess the risk of dropout has been recently implemented in North Carolina and finds that while logistic regression correctly classifies 40% of graduates while misclassifying eight percent of non-graduates, boosted decision trees increases the correct classification of graduates by approximately 20% at the same misclassification rate.[14]

The importance of both identifying at-risk students and creating actionable information is underscored by some states and localities already employing early warning indicator systems using longitudinal data analysis. In 2005, South Carolina began development of a longitudinal database, containing students as far back as 3rd grade and following them into high school, to be used by school personal such as counselors and administrators responsible for local at-risk models. In 2006, Maine implemented a K-12 integrated data system allowing for the assessment of likelihood of dropout using 9th grade indicators, and by 2012 started revising student data collection to expand early education indicators. In 2010, Massachusetts began developing an Early Warning Indicator System leveraging P-12 data to predict proximate outcomes, such as the likelihood of reading proficiency in 3rd grade and passing all 9th grade courses, and soon after released a state-wide Early Warning Indicator Index using academic and behavioral student characteristics[15] to identify drop-out propensity by risk level of first-time 9th grade students in large urban districts (Curtin et al., 2012). The steady release of such systems communicates a necessity for a robust prediction method that not only identifies whether students are at-risk but communicates actionable information to their stakeholders.

Most closely related to the work we describe here is a working paper by Austin et. al (2020) which uses administrative data from six states to study the extent to which a student's rank in the distribution of academic performance changes during their schooling career. Using test score data from 3rd grade, the authors predict percentile rank of student test scores in 8th and 10th grades and high school graduation. They focus on the extent to which there is variation between districts in *academic mobility*, i.e., movement of students in the test or graduation probability distribution since 3rd grade. While they find there is significant heterogeneity across districts in academic mobility, 3rd grade test scores are highly predictive of students' positions in high school test and graduation probabilities in all states.[16]

The growing body of research clearly shows increasing interest in predicting long-term student outcomes, how early measures of student characteristics and achievement are associated with these outcomes, and finally, providing actionable information for individual student improvement. There are limitations, however, to the studies described above. None, for instance, examine the degree to which early test scores predict advanced course-taking, the degree to

[14] See also Christie et al. (2019), which uses gradient-boosted decision trees to predict the risk of dropout.

[15] Student characteristics include spring 2011 8th grade MCAS results, spring 2011 8th grade English Language Arts (ELA) scores, 2010-11 attendance rates, number of suspensions in the 2009-10 and 2010-11 school years, and age as of September 1st, 2011.

[16] For instance, the coefficient indicating the relationship between a student's position in the 3rd grade test distribution and the 10th grade English Language Arts test distribution is the neighborhood of 0.8.

which segments of test distribution information (e.g., 3rd to 8th grade and then 8th grade to high school graduation) may be pieced together to make long-term student outcome predictions, and the degree to which estimates of relationships from one state provide accurate assessments of outcomes for students. We use the data described in the following section to focus on these issues.

3. Data Sources, Sample Inclusion, and Measures

To assess the predictive capacity of early-education student characteristics on long-term outcomes, we use longitudinal student panels from Massachusetts, North Carolina, and Washington, including student characteristics and test scores from 3rd grade up to 12th grade between 1998 and 2018 which, depending on the outcome and state, contain as many as 11 cohorts of students.[17] The data across the three states are similar in that for each state we have measures of historical test scores, student characteristics, and three long-term high school outcomes: test scores in high school, advanced course-taking in high school, and high school graduation.

The Massachusetts longitudinal student data combines annually reported test scores from the Massachusetts Comprehensive Assessment System (MCAS), course membership information from the Student Course Schedule, and demographic information and high school exit codes from the Student Information and Management System, all of which are provided by the Massachusetts Department of Education. The North Carolina longitudinal student data combines annual North Carolina Education Research Data Center End-of-Grade files, Masterbuild files, and AP course membership files, which include student-level characteristics such as URM and EDS status,[18] AP course taking behavior, and high school exit codes. The longitudinal student data in Washington combines the state's Core Student Records System and Comprehensive Education Data and Research System both maintained by the Office of the Superintendent of Public Instruction, which detail student-level characteristics such as URM and EDS status and high school exit codes as well as AP course membership.

This panel allows us to leverage students' standardized test scores throughout their academic career and link these scores to their high school course-taking patterns and graduation. There are, however, reasons to be cautious about generalizing our findings. The first is that the high school outcomes we investigate, while similar, may vary across states. Indeed, even in the three states we utilize, they are somewhat inconsistent. For instance, in the case of test scores, the majority of students in Massachusetts take a high school math test in the 10th grade, whereas the majority of students taking a math test in North Carolina and Washington varies by grade depending on the year. We circumvent this by limiting high school math test sample to cohorts with standardized testing regimes across students and calculate test score percentiles by grade, year, state, and test

[17] We observe partial cohorts of students for years following 2013 due to earlier-than-expected positive outcomes but do not include them in our analysis. For example, we observe 1,016 3rd grade students in 2008-2009 graduating prior to the end of 12th grade, but since we do not have access to data in 2018-2019, we do not see any students who would have graduated at a normal pace.

[18] Recall that EDS is an indicator that a student qualifies for free- or reduced-price meals.

type.[19] Similarly, standards for what constitutes advanced course taking or high school graduation may also vary across states (we discuss this more below).

A second concern is that, given the nature of the study, it is necessary to restrict our sample to students who were enrolled in public schools in the 3rd grade and have at least one public school high school outcome, also in a public school. Students may exit the sample by transferring to private or homeschool within the state, transferring out of the state, enrolling in-state but not attend school in the following year, or otherwise having an unknown exit status.[20]

In **Figure 1**, we show the average percentage of 3rd grade cohorts who are observed in subsequent grades by grade and by state to provide a sense of the nature of attrition. While it is possible that 3rd grade students in the sample may leave the sample and return in a later grade, the lines are monotonically decreasing across all three states. Attrition between 3rd grade and high school is between 15-30% of students in the three states but is considerably higher in North Carolina than in Massachusetts and Washington, where the average attrition from 3rd grade cohorts is quite consistent from grade-to-grade. It may be that these differences are attributable to features of the states' educational landscape such as trend in private school enrollment or compulsory education laws,[21] but detailed exploration of differences in sample attrition across states is outside the scope of this study.

Regardless of the reason for it, sample attrition is potentially problematic in assessing predictive power of early academic measures on high school outcomes for an average 3rd grader. For example, predictions will be biased if there are *unobserved* attributes of students who leave the samples that are correlated with the likelihood of out-of-state mobility, 3rd grade test scores, and high school achievement. While we are unaware of any direct evidence on this issue, there is ample evidence on the role of parental involvement on student outcomes (e.g., Castro et al., 2015; Henderson, 1994; Wilder, 2014) and that low-income families, who also tend to have lower achieving students (Reardon, 2011), are more likely to be mobile (Mehana & Reynolds, 1995, 2004). Thus, it is no great leap to imagine that unobserved attributes are correlated with mobility and achievement.[22] Such a relationship would lead to bias in the estimated parameters of the model (we describe how we address the potential issues associated with missingness bias in Section 4.1).

[19] All cohorts in Massachusetts take a standardized math test in 10th grade, only a single observable cohort in our sample Washington takes a standardized assessment in 11th grade, 83% students in North Carolina take an 11th grade test in 2006, and over 97% of students take an 10th grade test in 2008-2011. North Carolina has a transition year in 2007 where 93% of students take either the 10th or 11th grade equivalent math test.

[20] That is, we begin with the 3rd grade cohort in a state and year and follow that cohort longitudinally. We do this for all 3rd grade cohorts in each state and average the cohort retention results.

[21] For instance, there are also differences in compulsory education laws: under North Carolina's compulsory education laws, most students can legally drop out as soon as they turn 16, whereas in Massachusetts and Washington students must attend school until age 18. But the pattern of year-to-year attrition in **Figure 1** does not reflect a sharp divergence between states in high school, which is what one would expect were the differences in attrition to be related to compulsory education requirements. Though this is outside the scope of our analysis, it may be that there are state-to-state differences in private school enrollment or homeschooling.

[22] Parents of low-achieving students who tend to contribute more to their children's academic success (e.g. encouraging them to do homework) might, for instance, be expected to try to keep their children in a stable educational setting.

In addition to the restriction of the state samples associated with attrition, we also restrict our outcome measures for comparability across cohorts. Since we consider five-year graduation rate and advanced course-taking at any point in high school, in order for students in our analyses to be assigned a graduation or advanced course-taking outcome we require that they are either observed up to 12th grade, observed graduating within 5 years of entering high school, or observed dropping out, and only consider students where these conditions are possible to assess. Similarly, we only include students where these conditions can feasibly be observed for the entire cohort.

The number of cohorts we use for our analytic sample from each state varies according to the length of the administrative panel available. North Carolina has up to 11 cohorts and the largest samples per cohort, with the first cohort in third grade starting in the 1997-1998 school year. The average cohort for North Carolina is 67,460 for graduation outcomes, 66,840 using advanced course-taking outcomes, and 77,875 using high school testing outcomes. The samples in Massachusetts (three cohorts, with the first cohort enrolled in third grade in the 2006-2007 school year) and Washington[23] (four cohorts, with the first cohort enrolled in 3rd grade in the 2005-2006 school year) are similar with about 47,000 to 55,000 for the various outcomes. The specific sample sizes for each of the later high school outcomes we observe for the analytic sample are provided in **Table 1**. A notable discrepancy across states is in regard to high school math test outcomes, due to inconsistencies over time in standardized testing requirements. While Massachusetts has had a standardized state test continue throughout our panel, North Carolina's longest panel of comparable high school begins for students attending 3rd grade in 2006, and Washington did not fully phase a single standardized test requirement until 2018-19, corresponding to 3rd graders in 2010-11, the last observable year in our 3rd grade panel.

The table also provides selective descriptive statistics by state. Each state differs somewhat in their racial and socio-economic distribution, the biggest differences being a much larger proportion of African American students in North Carolina (27%) than the other two states (5% to 8%), and a notably smaller population of Asian and Pacific Islander students in North Carolina (two percent) than the other states (6% to 9%).

Of the 3rd graders that we track to high school, graduation is relatively similar across the states, in the range of 80-90% across individual cohorts.[24] Advanced course-taking varies significantly more: while 59% of students in the sample were found to take at least one advanced math or science course in Washington, and 69% in North Carolina, only 48% of students in Massachusetts are identified as taking an advanced math or science course. These differences are

[23] In 2009 there is a large decline in 3rd grade cohort size in Washington state for high school test score outcomes due to switching of high school testing schema over observed years, and approximately one third of Washington State schools participated in the state's Smarter Balanced Assessment pilot in the 2013–14 school year, so elementary test scores are not available in 2013–14 for students in these schools.

[24] We calculate an average graduation rate in Massachusetts of 91%, an average rate in North Carolina of 87%, and an average rate in Washington of 84% These calculated graduation rates are relatively similar to the recent state reports of high school graduation rates of 88% for Massachusetts in 2018 (according to the Massachusetts Department of Education, see http://www.doe.mass.edu/infoservices/reports/gradrates/), 86% for North Carolina in 2017 (NCES, 2020), and 79.3% for Washington in 2016 (Weaver-Randall & Ireland, 2018).

likely based on the definitions of the advanced course-taking measures.[25] For North Carolina and Washington, we use high school course names and a course taxonomy developed by Burkam et al. (2003) to identify advanced math and science courses, whereas in Massachusetts we use an indicator of advanced courses in combination with a subject area course code provided by the state (Massachusetts Department of Education, 2018); this indicator more narrowly defines what is an advanced course than is defined by Burkam et al.. Because we do not have course names in Massachusetts, we cannot directly assess the similarities of these courses to those in North Carolina and Washington, though as we show below (in Section 5), the findings for Massachusetts on advanced course-taking turn out to be similar to those for the other states.

Finally, we see similar patterns across the three states based on the quartile of 3rd grade math test achievement; lower scoring students tend to be disproportionately represented by various measures of student disadvantage (e.g., being EDS, or being in an URM category). And, not surprisingly, lower scoring students also tend to have substantially less positive high school outcomes.

4. Empirical Approach

Our empirical approach is designed to assess the accuracy of predicting high school outcomes based on 3rd grade test scores, identifying the early student characteristics most influential of the high school outcomes, and the amount of information lost in these predictions if we predict across states or in multiple stages. We look at three primary outcome measures: high school test scores in mathematics, advanced course-taking behavior, and graduation.

4.1 Analytic Approach

To assess the relationship between 3rd grade test scores and high school math score percentile M_i^{HS} we estimate both oft-used generalized linear models and more recently developed machine learning techniques for generating predictions. We begin by following the approach in Austin et al. (2020), which models the ranking of students in the high school math test distribution as a function of 3rd grade test ranking and observable student characteristics:[26]

$$M_i^{HS} = \beta_0 + T_i'\beta_1 + X_i'\beta_2 + \varsigma_i T_i'\beta_3 + \varsigma_i + \delta_i + \varepsilon_i, \quad (1)$$

Specifically, in (1), T_i is a vector of 3rd grade math and reading test score percentile categorized by subject for student i, X_i is a vector of student $i's$ characteristics including race, gender, disability status, English language learner (ELL) status, EDS status, and enrollment status in special education, ς_i is a state fixed effect, δ_i is a year fixed effect, $\varsigma_i T_i$ represents a state-test score percentile interaction, and ε_i is a mean-zero error. We interact the state and year fixed effects with the vector of 3rd grade math and reading test score percentile. Our primary focus is on β_1, which indicates the relationship between 3rd grade tests and high school outcomes (high

[25] As a robustness check, we code an alternative definition of advanced course-taking in MA by following the course taxonomy of Burkam et al. (2003). We find a correlation of .76 in advanced course between the two definitions, and a correlation of .86 in the resulting predictions.

[26] We discuss how sample attrition may affect estimates and model prediction in Section 5.3. As we describe in that section, sample attrition does have much impact on model predictions.

school math tests in (1)), which is recoverable from the marginal effects of the math and reading test scores.[27] We also estimate several extensions of the above model, including replacing or supplementing 3rd grade test rankings with 8th grade rankings on math and English Language Arts (ELA) tests, estimating the various specifications separately by state, and estimating models with 8th grade math score percentile as an outcome.[28]

For each binary outcome, whether students take advance courses and graduate from high school, we estimate a conditional probit model defined by:

$$P(Y_i = 1 \,|T_i,\ X_i) = \ \Phi(Z_i), \tag{2a}$$

$$Z_i = \ \beta_0 + \ T_i'\beta_1 + X_i'\beta_2 + \varsigma_i T_i'\beta_3 + \ \varsigma_i + \ \delta_i + \varepsilon_i, \tag{2b}$$

where Y_i is the outcome, $\Phi(\cdot)$ is the cumulative normal distribution, and T_i, X_i, ς_i, δ_i are consistent with (1) above, and we also assume that ε_i is a mean-zero error term. And, as above, we estimate specifications of (2) which either replace or supplement 3rd grade test rankings with 8th grade rankings on math and ELA tests, and estimate these specifications separately by state. Furthermore, we estimate the above models on two additional outcomes: an indicator for scoring in the top half of the testing distribution for 8th grade math and high school math.[29]

While the R^2 and Pseudo-R^2 of the above specifications give some indication of model fit across different specifications, these are not necessarily representative of out-of-sample prediction accuracy. Thus, below (in Section 4.2) we define metrics to allow us to assess the accuracy of out-of-sample predictions that, in particular: 1) are based on the classification of students into categories; 2) allows us to assess the extent to which using earlier (3rd grade) test score information leads to different predictions than later (8th grade) test information, and how the omission of one diminishes the fit of the model; and 3) compare the efficacy of using cross-state models and segments of achievement data (grades 3 to 8 then 8 to 10) to make long-term outcome predictions.[30]

There is evidence (Austin et al., 2020) that rank-rank relationship in test score outcomes is essentially linear in models such as those above, we are unaware of similar evidence when it comes to advanced course taking or high school graduation. Thus, to explore non-linearities and the possibility that the relationship between 3rd grade tests and high school outcomes is different based on students' 3rd grade background characteristics, we supplement the above models with

[27] Note that because test scores noisy measures of student learning $\beta_1$will be biased downward. It is possible to correct for this, as in Austin et al. (2020), using the standard errors of measurement (SEMs) associated with 3rd grade test scores across cohorts. We opt not to do this given that our interest is in estimating how well 3rd grade tests *predict* later achievement, so, for our purposes, it makes sense to use imperfect (noisy) test measures.

[28] The estimation of separate state models is important for testing the degree to which estimates from one state can reliability be used for generating predictions of student achievement in a different state. Results for this additional outcome is presented in **Appendix B Table B1**.

[29] Results for these additional outcomes are presented in **Appendix B Table B1**.

[30] In doing out-of-sample predictions, we omit year and cohort fixed effects.

more flexible specifications in which we include decile of test score achievement in 3rd grade and interactions between these deciles and student characteristics.

Relatedly, there is growing interest in the use of machine learning (ML) techniques for generating out-of-sample predictions. While these techniques are not widely used in education, they have been used in fields such as medicine (Kourou et al., 2015), genetics and genomics (Libbrecht & Noble, 2015), and time-series forecasting (Bontempi et al., 2012).

We use three ML approaches, described in greater detail in **Appendix A**, for predicting high school outcomes: 1) Kernel Support Vector Machines (kSVM), 2) Random Forest Classification (RFC), and 3) Gradient Boosted Decision Trees (GBDT). The advantage of these approaches over traditional generalized linear models is their flexibility in modeling our binary as nonlinear functions of the student characteristics. kSVMs model nonlinearities using a function of the distance between two points referred to as a "kernel". RFCs and GBDTs both improve prediction by iteratively splitting students along a single dimension of their observable characteristics according to what best separates graduates from non-graduates, or advanced course-takers from those that do not. This iterative splitting results in each student falling into a "bin" based on their observable characteristics, with the end goal that the majority of students in the same bin having the same outcome and allows for representation of complex nonlinear relationships. The difference is how the methods deal with overfitting while keeping high prediction accuracy. RFCs do this by averaging many classification models using random subsets of variables, whereas GBDTs increase accuracy by iteratively modeling the residuals generated by the prior classification tree.

While kSVMs tend to handle high dimensional data, avoid overfitting, and perform well when there is clear separation between classes, their accuracy depends largely on the arbitrary choice of an appropriate kernel, model estimation has a computational complexity cubic in the number observations (Abdiansah & Wardoyo, 2015), and they do not produce any interpretable probabilities of membership. RFCs and GBDTs are less computationally expensive to estimate but tend to overfit data with lower sample sizes (e.g., Bramer, 2007).

4.2 Measuring the Accuracy of Out-Of-Sample Predictions

A natural evaluation metric for predicting math test score percentile is root mean squared error (RMSE), as it is both in accordance with the minimization criterion for linear regression and penalizes larger deviations more heavily. But, how to compare the accuracy of student i's predicted probabilities $\hat{p}_i$ relative to dichotomous outcome variables is less straightforward. We focus on student i's high school graduation, which we will refer to as Y_i, without any loss of generality. Assignment of the model predicted probabilities to a positive (graduated, $Y_i = 1$) or negative (did not graduate, $Y_i = 0$) outcome depends on a threshold value c (i.e., we say a student graduates when $\hat{p}_i \geq c$). However, the categorization of whether students graduate depends crucially on the choice of c, and thus the number of Type I (students being predicted to graduate when they don't) and Type II (students being predicted not to graduate when they do) errors. As in Christie et al. (2019) and Geiser and Stanelices (2007),[31] we alleviate this issue by

[31] Geiser and Stanelices (2007) report the concordance rate, a small variation on the AUC which discounts points predicted to have equal success and failure probabilities.

reporting the Area Under the Curve (AUC) as an evaluation metric, which considers all choices of the threshold value, c, during model comparisons.

Define the True Positive Rate (TPR) and False Positive Rate (FPR) as

$$TPR(c) = \frac{\sum_i \mathbf{1}(\hat{p}_i \geq c, Y_i = 1)}{\sum_i P(Y_i = 1)}, \qquad FPR(c) = \frac{\sum_i \mathbf{1}(\hat{p}_i \geq c, Y_i = 0)}{\sum_i P(Y_i = 0)}, \tag{3}$$

where $\mathbf{1}(E) = 1$ if the event E is true and 0 otherwise. In other words, the FPR is the proportion of non-graduates that we misclassify as graduating, and the TPR is the proportion of graduates that we correctly classify as graduating.[32]

Letting $FPR(c) = x$, we define the Receiving Operator Characteristic (ROC) curve as a function over cut points defined as $f(c) = TPR(FPR^{-1}(x))$. This allows for a visualization of the classification accuracy of binary models across the entire range of $c \in [0,1]$. For example, a perfect model would yield $f(c) = 1$ for all values c, suggesting that all graduates are correctly classified no matter what threshold is chosen, including the threshold c^* that allows $FPR(c^*) = 0$. On the other hand, an uninformative model, i.e., equivalent to flipping a coin to make the prediction, would yield $TPR(c) = FPR(c)$, suggesting that the probability of correct classification and misclassification are equally likely.

To characterize the ROC curve across all values of the threshold c, we use an evaluation metric called the AUC, defined as

$$AUC = \int_{c=0}^{1} TPR(c) dFPR(c). \tag{4}$$

Rather than representing a predictive model using a single threshold value, equation (4) provides an overall metric for model performance across all thresholds comparable across predictive models, where an AUC of 1 corresponds to a perfect classification model, and an AUC of ½ corresponds to essentially random classification.

We report traditional goodness-of-fit measures for all models, but we are also interested in assessing the out-of-sample prediction accuracy. This is important because unobserved factors correlated with cohorts could lead to biased estimates of how well early measures of academic achievement predict later outcomes.[33] Additionally, there is good evidence that evaluation of prediction accuracy metrics on the sample in which a model is estimated tend to be overly optimistic, as the model is specifically catered to minimize the sample error (Picard & Cook, 1984). And, moreover, a primary reason for assessing the predictive power of elementary tests is to assess the extent to which they might be relied on for school-based early warning systems,

[32] In context of early warning systems, for instance, one might be especially worried about the number of at-risk students mistakenly labeled as future graduates. In this case, one might set the threshold c to be relatively high. However, this has a direct impact on the TPR—the number of future graduates we correctly classify. The choice of the threshold c represents a context-specific tradeoff between the FPR and TPR, which makes model comparison at an arbitrary value of c dangerous to generalize.

[33] Shores and Steinberg (2017), for instance, find that the economic shock of the Great Recession negatively affected student achievement.

such as those described in Section 2, and/or to provide parents with objective information about their students' likely educational trajectories (e.g., Learning Heroes, 2018).

To validate general out-of-sample prediction accuracy we use 10-fold cross validation (Hastie et al., 2005). Specifically, we are interested in estimating the expected *test error* TE:

$$E[TE] = E[L(\hat{p}(X), Y)], \tag{5}$$

where Y is a random variable associated with an outcome, X is a vector of random covariates, $L(\cdot,\cdot)$ is an evaluation metric and $\hat{p}$ is the estimated relationship between Y and X. Specifically, without loss of generality, let $Y = M^{HS}$ be a student's high school math test, X be a set of 3rd grade characteristics, and $L(\hat{p}(X), Y) = \frac{1}{n}\sum_i(\widehat{M^{HS}} - M^{HS})^2$ be the mean squared error between the predicted and actual high school math values. Using a set of observed data $(\tilde{X}, \tilde{Y})$ directly to calculate (4) can give an overly optimistic notion of our predictive accuracy if the optimal parameters of the evaluation metric $L(\cdot,\cdot)$ coincide with the parameters estimated to best fit $\hat{p}$; since estimates $\widehat{M^{HS}}$ are obtained by minimizing $L(\tilde{X}, \tilde{Y})$, we would expect this value to be smaller than if we were to calculate the squared error on a new dataset. One method of circumventing this issue is to randomly split the observations into ten equal partitions $K_1, \dots, K_{10}$ (or, more generally, κ partitions) and calculate the expected test error:

$$\widehat{TE} = \frac{1}{10}\sum_{l=1}^{10} L\big(\hat{p}_{K_l}(\tilde{X}_l), \widetilde{Y_l}\big), \qquad l = 1, \dots, 10 \tag{6}$$

where $(\tilde{X}_l, \widetilde{Y_l})$ is the subset of $(\tilde{X}, \tilde{Y})$ in partition K_l and $\hat{p}_{K_l}$ is estimated on the set $\{K_i \mid i \neq l\}$. In the context of high school test scores, equation (6) refers to the average mean squared error across 10 partitions. Estimates of the expected test error from equation (6) using 10-fold cross validation are upward-biased (Varma & Simon, 2006) and provide a conservative estimate of the true expected test error in equation (5).

The predictive validity of out of state models using this procedure may be influenced by the relative sample sizes of the state. For example, an out-of-state model may have better out of sample predictive accuracy than an in-state model by consistently observing a more complete sample of the student population. To deal with this issue, we propose a modification of the above cross-validation procedure. For each state $S \in \{MA, NC, WA\}$ we randomly split the sample into 10 equal partitions, $K_1^S, \dots, K_{10}^S$, and for each partition index we calculate the smallest number of observations m_i across all states:

$$m_i = \min\big(|K_i^{MA}|, |K_i^{NC}|, |K_i^{WA}|\big), \tag{7}$$

where $|\cdot|$ represents the number of observations in the partition. Then for each state S and each partition index i we randomly drop observations in each K_i^S until $K_i^S = m_i$. Finally, for each state pair (S_1, S_2), we calculate the expected cross-state test error:

$$\widehat{CTE}(S_1, S_2) = \frac{1}{10}\sum_{l=1}^{10} L(\hat{p}_{K_l^{S_2}}\big(\tilde{X}_l^{S_1}\big), \tilde{Y}_l^{S_1}), \tag{8}$$

where $(\tilde{X}_l^S, \tilde{Y}_l^S)$ is the subset of data from state S in partition l. When $S_1 = S_2$, this becomes estimated within-state test error represented by equation (6). Since randomly dropping student observations from the largest states in equation (7) may lead to substantially higher variation in the cross-state test error, we repeat the 10-fold cross validation procedure 100 times to ensure a representative description of cross-state test error is obtained for each state pair.

In addition to comparing in-state predictions to out of state estimates, we also provide direct correlations between the predictions. However, both of these aggregate measures may mask where in the predictive distributions the estimates diverge. To get a sense the degree to which the different use of parameters influences predictions throughout the predictive distribution, we follow Goldhaber et al. (2019) and first estimate equations (1) and (2) using students in Washington, Massachusetts, and North Carolina separately. Then, letting I_S be the set of all students who attend state S, for each distinct state pair (S_1, S_2) we estimate the in-state predictions from S_1 with the cross-state predictions from S_2 using the cubic polynomial model:

$$\hat{p}_{i,S_1} = \alpha_0 + \alpha_1 \hat{p}_{i,S_2} + \alpha_2 \hat{p}_{i,S_2}^2 + \alpha_3 \hat{p}_{i,S_2}^3 + \varepsilon_i, \tag{9}$$

where $\hat{p}_{i,S_j}$ is the predicted probability of an outcome for student $i \in I_{S_1}$ using model coefficients calculated on students in I_{S_2}. This allows us to see particular aspects of prediction model differences. Differences between the estimated mean trend on the right-hand side of equation (9) and within-state predicted probabilities $\hat{p}_{i,S_1}$ represent model variation across different magnitudes of outcome probability. For example, since Washington has a lower graduation rate than Massachusetts, estimates from Massachusetts may distinguish students with lower likelihoods of graduation more poorly than in-state estimates.

We follow a similar procedure to assess the use of segments of test score data to make long-term projections. We assess this issue because the length of administrative panels in some states are limited such that it may not be possible to predict high school outcomes based on 3rd grade test scores (Data Quality Campaign, 2016). Thus, it is not possible to predict outcomes, like high school graduation, based on 3rd grade test scores. But one could use the parameters from models predicting test relationships between third and 8th grade (segment 1) and 8th grade tests to high school graduation (segment 2) to predict across the two segments so as to link 3rd grade tests to high school graduation.

Let the superscripts ρ_i denote distinct panels of students. We first estimate the relationship between third and 8th grade student observables by equation (10a) on a panel of students ρ_1, estimate the relationship between 8th grade student observables and long-term outcomes by equation (10b) on a distinct panel of students ρ_2, and with the resulting estimates predict long term outcomes using equation (11):

$$Z_i^{p_1} = f(\alpha_0 + T_i^{\rho_1 \prime}\alpha_1 + X_i^{\rho_1 \prime}\alpha_2 + \gamma_i), \tag{10a}$$

$$Y_i^{p_2} = g\left(\beta_0 + Z_i^{\rho_2}\beta_1 + \eta_i\right), \tag{10b}$$

$$\widehat{Y_\iota} = g\left(\widehat{\beta_0} + \widehat{Z_\iota}\,\widehat{\beta_1} + \varepsilon_i\right),$$

$$= g(\widehat{\beta_0} + f(\widehat{\alpha_0} + T_i' \, \widehat{\alpha_1} + X_i' \, \widehat{\alpha_2})\widehat{\beta_1} + \varepsilon_i), \tag{11}$$

where X_i is a vector of 3rd grade student characteristics, T_i is a vector of student test score percentiles, Z_i is a vector of 8th grade student characteristics and outcomes, $f(\cdot)$ and $g(\cdot)$ allow representation of either linear or probit regression models, and $\gamma_i, \ \eta_i, \varepsilon_i$ are mean-zero errors. While **Appendix A** shows we might expect some loss in accuracy, we will show in the next section that estimation using this two-stage procedure does not dramatically reduce predictive accuracy.

The capacity for out-of-sample generalizability may be affected both by differences in educational landscapes across states and by differences in non-overlapping time periods. For example, when using a model trained on graduation probabilities for students in Massachusetts from 2007-09 to predict graduation probabilities for students in North Carolina from 1998-2008, the relationship between student characteristics and graduation may be different between Massachusetts and North Carolina but may have also changed over the course of this decade. We test this by using a nested model likelihood ratio test comparing estimates from equations (1) and (2) with equivalent models fully interacting year for each state, and overall.[34]

5. Results

5.1 Baseline Findings on Long-Term Predictions

Our main findings are reported in **Table 2**, which shows the relationship between 3rd, 8th, or both 3rd and 8th grade tests, other student-level covariates (in the 3rd grade) and three high school outcomes: percentile on a high school math test, the probability of taking an advanced math course, and the probability of graduating.[35] While not reported in the table, the models also include state and cohort indicators and interactions between these and base year test scores. These interaction terms are statistically significant for all outcomes, suggesting that the relationships between base year test scores and outcomes differ by state and cohort.[36]

Columns 1-3 report on models estimating high school math test scores for specifications that include: 3rd grade scores (Column 1); 8th grade test scores (Column 2); and both 3rd and 8th grade test scores (Column 3). Not surprisingly, prior test scores are highly predictive of high school scores. This is true for both math and reading when the tests are entered into the models independently, with 3rd grade math and reading scores being smaller in magnitude than 8th grade scores. But we also see (in Column 3) that both 3rd and 8th grade tests are significant when

[34] Since the model likelihood ratio test does not extend to probit regression, we use a linear probability models on binary outcomes. We omit high school math tests in Washington state from this robustness check since there is only a single year in the panel.

[35] The high school math test score models are estimated by OLS so the coefficients on base year test scores represent the estimated effect of a change in base year test percentile on the change in high school test percentile. The course-taking and high school graduation probabilities are estimated by probits, but we have converted the coefficients into marginal effects, so, for instance, they show how a change in a 3rd grade test percentile is estimated to affect the probability of high school graduation.

[36] As we discuss below, while statistically significant, the differences between these relationships across states and over time are arguably of minor practical significance.

included simultaneously, suggesting that the trajectory of achievement matters for high school grade predictions (more on this point below).[37]

In Columns 4-6 and 7-9 of Table 2 we present analogous specifications for the probability of advanced course-taking and high school graduation, respectively. The trends in the coefficients on base year test score are similar to those described above for high school tests, though the association between base year test scores and these outcomes is smaller. This is especially true for the relationship with high school graduation, which is not surprising since the large majority of students in the sample do graduate. In particular both third and eighth grade tests are statistically significant in the same model for both advanced course-taking (column 6) and high school graduation (column 9). This suggests both are important in making predictions of these outcomes, however, as we show below (in Section 5.2), it turns out that the predictions about high school outcomes are little affected by whether they are generated using third or eighth grade tests.

Consistent with prior evidence (e.g., Austin et al., 2020; Hernandez 2011; Zau & Betts, 2008), there is a strong correlation between early measures of test achievement and later high school outcomes. In particular, we show in Table 4 quite strong correlations between a student's place in the distributions of 3rd grade and high school math tests (column 1), and 8th grade and high school tests (column 2).

We assess the importance of trajectory by estimating variants of the above models where we include math and reading test scores for all permutations of 3rd, 4th, and 5th grade. State and year effects are omitted for the purpose of valid out-of-sample prediction. We examine the relationship between outcome accuracy measures and the combination of test scores. **Table A3** shows out of sample accuracy metrics for high school math tests, advanced course-taking, and graduation. We note that outcomes are differentially affected by the testing grade. For example, the root mean squared error is negatively correlated with the grade level, whereas advanced course taking and graduation do not have such a correlation. Secondly, confirming recent evidence (Fazlul et al., 2021) that there is a diminishment in information due to gaps in student test data, we see an increase of .02 in predictive accuracy as measured by AUC when considering all three grades at once for advanced course-taking relative to only 3rd grade.[38]

The trajectory of a student's elementary test scores also matters for estimates of future achievement. As a practical example of these effects, the estimated probability of a student scoring in the 30th, 40th, then 50th percentiles in math in 3rd, 4th, and 5th grade, respectively, is 85%, whereas the estimated probability of a student scoring in the 50th, 40th, then 30th percentiles in math in 3rd, 4th, and 5th is 83%. Put another way, the risk of not graduating is 13% larger for the student with declining percentile ranking. Trajectory similarly influences advanced

[37] Note that 3rd grade reading test scores are actually negative in this model. This is not terribly surprising given the strong correlation between 3rd grade math and reading test scores (0.72); when we re-estimate column (3) using single subject test scores for both grades, both grades are positive and statistically significant.

[38] Note that Fazlul et al. focus on school and district performance ratings rather than predictions of individual student achievement. Testing gaps induce additional issues in estimating school or district performance. For instance, assuming that testing starts in the 3rd grade, many schools would have no students with pre- and post-test scores for their time in a particular school building (e.g., K-5 school) if test scores are missing for two consecutive years.

course taking and high school math percentile; the student with a declining trajectory has a likelihood of 53% of participating in advanced course-taking and is predicted to score in the 40th percentile in high school math, whereas the student with an increasing trajectory has a likelihood of 60% to participate in advanced course-taking and predicted to score in the 45th percentile of high school math.

Also consistent with prior evidence, holding constant base year test achievement, there remain significant differences in test achievement associated with a student's 3rd grade characteristics, with traditionally disadvantaged students (Hispanic and African American, those receiving EDS, etc.) predicted to be significantly lower in the high school test distribution. Relative to non-EDS students, for instance, EDS students are predicted to be two to six percentile points lower in the high school test distribution.

In Table 3, we estimate models on each of the three states separately. For each state and outcome both math and reading 3rd grade test scores are strongly predictive of high school outcomes. We also estimate more flexible specifications by estimating each model separately by state, allowing all relationships to differ by state, specifying deciles of 3rd grade math and reading achievement to allow for nonlinearities in the relationship between 3rd grade achievement and high school outcomes, and by interacting those deciles with each student's 3rd grade EDS classification to assess whether the relationships between test scores and high school outcomes differ based on their income status.

Chow tests show that the fit is better for models that are estimated on each state separately. Students' positions in the 3rd grade math distribution are similarly predictive of high school graduation to their positions in the 3rd grade reading distribution. Not surprisingly, however, 3rd grade math percentiles are much more strongly predictive than reading percentiles of advanced course-taking *in math and science* and high school *math* test percentiles.[39] All else equal, a student at the 10th percentile of the 3rd grade math test distribution rather than the 90th percentile is expected to be 38-42 (depending on state) percentile points lower in the high school math test distribution, is expected to be 35-41% less likely to take an advanced course in high school, and 12-14% less likely to graduate.

Figures 2 to **4** show marginal effects of high school outcome by EDS status and test score decile. [40] Specifically, by each decile of math or reading and by EDS status, we plot mean probability of graduation, probability of advanced course-taking, and high school math test percentile, along with 95% confidence intervals. Visual inspection of the figures shows a linear relationship between 3rd grade test percentile and each of the outcomes, with the exception of the graduation probability at the tail of the test score distribution; students scoring in the lowest

[39] These results hold when predicting 8th grade math percentiles, see **Appendix B Table B1**.

[40] As noted above, there are also differences by other characteristics, such as student race/ethnicity, but we highlight the differences by EDS status because they tend to be much larger than those for the other student sub-groups.

decile in math and reading are about 10% less likely to graduate than students in the second decile—an effect three times larger than differences across the other adjacent deciles.[41,42]

These figures also highlight consistent, large effects of EDS status on predictions. Specifically, within test score decile EDS students are 8percent to 10% less likely to graduate across all test scores for math and reading—approximately the same effect as going from the 2nd to the top decile of scores. Furthermore, EDS recipients are 7 to 12% less likely to take an advanced course depending on test score decile and positively correlated with test score, lowering probability of taking an advanced course in high school as little as a few percent at the bottom of the testing distribution, up to eight percent at the top of the distribution. Finally, EDS students consistently score three to five percentile points lower on the high school math distribution.[43,44]

5.2 Out-of-Sample Prediction Accuracy

Figures 5 and **6** present ROC curves and related AUC on the entire sample related to predicting graduation and advanced course-taking. There is a striking similarity in the precisions using 3rd grade student tests, AUC of 0.755, versus 8th grade student tests, AUC of 0.782. Indeed, the finding that 3rd grade tests predict high school outcomes nearly as well as 8th grade tests might be interpreted as suggesting the need to intervene earlier as these findings are consistent with evidence that student achievement gaps form early (von Hipple et al., 2018; Zau and Betts, 2008) and the trajectory of achievement is little affected during schooling (Austin et al., 2020).

The difference is approximately 40% larger for advanced course-taking, with an AUC of 0.799 for 3rd and 0.836 for 8th grade, but this is not terribly surprising, since high scores in 8th grade ELA may result in direct sorting into an advanced course in ninth grade. Furthermore, there is still relatively high predictive power on high school course-taking behavior in 3rd grade, suggesting the ability for school systems to conduct earlier and potentially more impactful interventions and communicate information to parents about student trajectories for a multitude of academic goal posts.

While AUC provides such a standardized quantification of model performance by aggregating over all cut points, the measure is not terribly intuitive. Some authors such as Allensworth and Easton (2007) and Sorenson (2019) report model performance at specific cut points. To facilitate comparison with their work, we also report our predictive accuracies using these cut points. With a wide array of 9th grade characteristics, such as GPA, credit completion

[41] These patterns hold at the individual state level and using 8th grade math percentile as an outcome. Marginal effects by state are presented in **Appendix A Figures A22-A32**.

[42] Large effects of EDS are also present when predicting 8th grade math percentile, probability of scoring in the top half of the 8th grade testing distribution, and probability of scoring in the top half of the high school testing distribution. See **Appendix A Figures A4-A6**.

[43] We test the statistical significance of the differential nonlinear relationship by EDS status by interacting EDS status with a quadratic polynomial of test score and find that the coefficients are highly significant ($p < 0.001$).

[44] Differential effects based on non-academic characteristics are not limited to EDS status. We find that on average, some racial and ethnic groups have a different probability of achieving high school outcomes, all else equal. For example, American Indians are five percent less likely than white students to graduate, all else equal, and three percent less likely to take advanced courses in high school.

and number of course failures, Allensworth and Easton (2007) correctly classify 85% of graduates while misclassifying 28% of non-graduates. In comparison, our models relying on 3rd grade tests and student characteristics correctly classify 65% of graduates while misclassifying 28% of non-graduates.[45] While this difference may seem substantial, characteristics such as course-failures in 9th grade are undoubtedly impactful in ensuring high school students graduate within four years. Furthermore, Allensworth and Easton (2007) predict on-time graduation, which tends to have a stronger relationship with early test scores (Austin et al., 2020). Predicting graduation in North Carolina, logistic regression models reported by Sorenson (2019) and our models both correctly classify 40% of graduates while misclassifying eight percent of non-graduates.

Figures 7 to **9** show 10-fold cross validated AUC predicting graduation and advanced course-taking, and RMSE predicting high school math tests (as well as 8th grade tests) using within-state estimates, cross-state estimates and two-stage estimates.[46] Not surprisingly, the within-state estimates (i.e., those that use parameters based on models spanning 3rd grade through high school) have consistently higher prediction accuracy and consistently lower RMSE than cross-state and segmented estimates. However, the information benefit of using within state parameters is limited and often statistically insignificant. For instance, in both Massachusetts and Washington the within-state estimates for high school graduation are statistically indistinguishable from segmented estimates.[47,48]

In some cases segmented estimates are more accurate than the cross-state estimates. In the case of both high school graduation and advanced course-taking, segmented estimates from Washington and Massachusetts produce statistically significantly higher AUCs than the cross-state estimates. When comparing cross-state estimates and segmented estimates for high school math scores, segmented estimates perform statistically significantly better than half of the cross-state estimates. However, while statistically significant, many of these differences are inconsequential from a practical standpoint, suggesting that the bias associated with estimating models from separate student panels is comparable to the bias associated with differences in state education systems.[49]

The high correlation between model predictions across states, ranging from .835 to .997 and shown in **Table 5** suggests a large amount of agreement in the relationship between 3rd grade

[45] When using 8th grade tests, we classify 70% of high school graduates while misclassifying only 28% of non-graduates.

[46] Point estimates and 95% confidence intervals for each metric are generated by the observed mean and quantiles of the 10,000 values produced by 100 repetitions of 10-fold cross validation for each outcome (Vanwinckelen & Blockeel, 2012).

[47] Due to the potential opacity of the AUC estimate, we also show accompanying ROC curves across model specification for graduation in each state in Appendix **Figures A19-A21**. These curves illustrate the striking similarity between model specifications across the entire threshold distribution.

[48] Similar patterns are seen when predicting 8th grade test percentile, probability of scoring in the top half of the 8th grade math testing distribution, and the probability of scoring in the top half of the high school math testing distribution. See Appendix **Figures A10-A12**.

[49] The segmented parameter models assume students are missing completely at random (MCAR) from their cohorts, which may not be the case in practice.

student characteristics and high school outcomes.[50] However, it is possible that this correlation masks divergences in the predictions at different points in the prediction distribution. Thus, **Figures 10-18** also show pair-plots comparing in-state and cross-state predicted probabilities along with the estimated polynomial regression line represented in equation (9). The linear relationship between predicted probabilities of graduation between Massachusetts and North Carolina suggest that the models share similar information across the probability distribution. However, the nonlinear mean trend between both states when compared to Washington, particularly in the lower tail of the probability distribution, suggests noisier estimation of the most at-risk students. Similarly, the distinct relationship and large variance in predicted probabilities for advanced course-taking between North Carolina and the other states suggests a poor match between student patterns in advanced course-taking behavior based on 3rd grade student characteristics. Finally, as is apparent from **Figures 16-18**, there is linear relationship between predicted probabilities estimated from in-state and cross-state models, suggesting a consistently estimated relationship between student test scores across the high school math testing distribution.[51]

5.3 Assessing the Potential Implications of Sample Attrition

As briefly discussed above (see discussion in Section 3), there is concern that sample attrition could lead to biased estimates of the relationship between 3rd grade test achievement and high school outcomes, i.e., the estimates for students who remain in the sample may not reflect the relationships for the entire sample of students that are first observed in the 3rd grade given that students who leave the sample may have a different relationship between their early test scores and long-term outcomes than those who stay. Indeed, when we regress sample attrition on 3rd grade characteristics and decile of test score, we observe a significant negative relationship between test score and missing high school outcomes. The likelihood of observation through 12th grade for students in the lowest decile of both math and reading achievement, for instance, is about 13 percentage points lower than those scoring in the middle of the testing distribution,[52] and EDS students are about 7 percentage points less likely to be observed into 12th grade than their non-EDS peers. These findings are consistent with recent evidence that mobility is a measure of students being at-risk (Goldhaber et al., 2021). We address the issue in two ways, both of which suggest that sample attrition has little impact on the model estimates.

First, we assess model coefficients and predicted outcomes with subsets of the data that are defined based on missingness at different points in students' academic careers. In particular, while high school outcomes are missing for some students with unknown exit behavior, we are able to see many of their 8th grade test scores. Hence, we can get a sense of potential missing data bias by comparing the relationship between early academics for students who have

[50] Since North Carolina has a substantially longer panel than other states, exhibiting large changes in yearly rates of graduation and advanced course-taking, we adjust predicted probabilities for these year effects.

[51] Estimates travel well across state lines throughout the prediction distribution when predicting 8th grade test percentile, probability of scoring in the top half of the 8th grade math testing distribution, and the probability of scoring in the top half of the high school math testing distribution. See Appendix **Figures A7-A9**.

[52] This relationship also holds for each state individually.

observable outcomes in high school and those whose highest-grade scores are in 8th grade.[53] We conduct regression analysis of 8th grade test scores on 3rd grade test scores with and without students with missing high school outcomes. The magnitude and direction of the difference in the resulting coefficients provide intuition on the severity and direction of potential bias. [54] The difference in the magnitude of the relationship between 3rd grade math test percentile and 8th grade math test percentile is .03 between the unrestricted and restricted samples, suggesting relatively little influence due to sample composition. Similarly, the difference between 3rd grade reading test percentile and 8th grade reading test percentile is .006, where these small differences hold when analyzing across states. Since we are mainly interested in whether the predictions change, we also use the estimates from the unrestricted and restricted samples to generate predictions of students' placement in the 8th grade test distribution. The correlation between the restricted and unrestricted samples across all states are greater than 0.99.

Second, we impute outcomes using ad hoc adjustments (see Austin et al., 2020) to test score effects for students with missing high school outcome data to bound the potential bias arising from a differential relationship between early test scores and high school outcomes. Specifically, for each outcome we estimate five variations of models (1) and (2): using student test percentile ranks from 3rd through 8th grade, 3rd through 7th grade, 3rd through 6th grade, 3rd through 5th grade, and 3rd through 4th grade.[55] We then generate imputed values for each outcome using the most informative model available for exiters.[56] We generate baseline imputations as well as ad hoc imputations designed to bound the potential bias. In particular, consistent with Austin et al. (2020), we assume that the relationship between 3rd grade tests and high school outcomes is increased or decreased by 10% and 25%. We then re-estimate models (1) and (2) including students with both imputed and non-imputed outcomes and compare the coefficients with our main results in **Appendix B Tables B2-B5**. Differences between columns (1) and (2) in these tables point to a difference in the distribution of characteristics between students with observable and non-observable high school outcomes, whereas differences between columns (2) and (3)-(6) represent the potential effect of a different relationship between early test score and high school outcome. Regardless of the ad hoc adjustment level, coefficients on 3rd grade math and reading test score percentiles deviate from the observable sample by less than .02. This striking similarity of coefficients across all outcomes and effect adjustments shows that the missingness present in the data is not likely to be a significant source of bias in the estimated relationships.

5.4 Comparison with Machine Learning Methods

[53] This is contingent on the underlying attrition process being similar in 3rd to 8th grade as 9th to 12th grade but exit reason after 8th grade may have differential causes. However, exits due to private school, for example, have similar rates before and after 8th grade.

[54] It does not, however, provide information about the effect of attrition between third and 8th grade, or the effect of attrition on the predictive power of 3rd grade tests on 10th grade test achievement. However, it is reassuring that the findings described below are similar if we instead focus on 6th grade test scores as the dependent variable.

[55] Since using this imputation procedure with probit models would require an arbitrary choice of cut point, we estimate linear probability models for the binary outcomes.

[56] For each student, we use model estimates based on the longest contiguous span in which they are observed since 3rd grade. For the sake of imputation, students who exit the sample and return in the span of 3rd through 8th will be treated as if they permanently exited.

While we expect high school math tests to have a linear relationship with early student academic test scores (Austin et al., 2020), we examine the potential improvement in predicting graduation and advanced course-taking using from machine learning methods. We compare 5 models: the probit described in equation (2) of the main paper, “Interaction”, the model in equation (2) with the addition of interaction terms, “SVM”, the Support Vector Machine model described in **Appendix A**, the “Random Forest” model described in **Appendix A**, and Gradient Boosted Decision Trees described in **Appendix A**. In each model we control for 3rd, 8th, and both 3rd and 8th grade test scores in turn, and all models control for student characteristics including race, gender, disability status, English language learner (ELL) status, EDS status, and enrollment status in special education. We randomly partition the data into 20% to use as a testing set and train on the remaining 80%.

Table A1 in **Appendix A** shows the resulting AUCs for each of the methods. First, we find that there is predictive capacity in the interaction of early academic math and reading test score percentile. If we interact math and reading test scores, the models have slightly greater predictive accuracy and in the case of graduation, greater growth in accuracy as test scores are included. Second, we find that Random Forests does outperform the baseline model, with an average 5% increase in accuracy according to AUC. For graduation, this difference grows with additional test scores included in the model; including both third and eighth grade tests in the baseline probit model increases the AUC by .01, whereas including the same tests in the Random Forests model increases the AUC by .03. We believe these differences are primarily driven by nonlinearities, and due to the more complex models sacrificing interpretability, use them as our primary specification.

6. Conclusion

A large literature shows that early academic performance, measured primarily by test scores, is predictive of later academic success, and that there are significant gaps in student achievement by student disadvantaged status. Our findings reaffirm these findings. Indeed, across three states we find consistent and very strong relationships between 3rd grade test scores and high school tests, advanced course-taking, and graduation. For instance, all else equal, a student at the 10th versus the 90th percentile of the 3rd grade math test distribution is expected to be 38-42 (depending on state) percentile points lower in the high school math test distribution, is expected to be 35-41% less likely to take an advanced course in high school, and 12-14% less likely to graduate. We conclude that early student struggles on state tests are a credible warning signal for schools and systems that make the case for additional academic support in the near term, as opposed to assuming that additional years of instruction are likely to change a student’s trajectory. Educators and families should take 3rd grade test results seriously and respond accordingly; while they may not be determinative, they provide a strong indication of the path a student is on.

Consistent with a small body of evidence (e.g., Zau and Betts, 2008), we find limited differences in the predictive power of 8th grade over 3rd grade tests, suggesting that there is little change in the trajectory of student achievement after the 3rd grade. Specifically, information about 8th grade test achievement does add statistically significant explanatory power to models predicting high school outcomes, yet the additional information does not change predictions markedly. For instance, the correlation in graduation, advanced course-taking, and high school math predictions between models using 3rd grade test scores and models using 8th grade test scores are .86, .94, and .82 respectively.

Our results illustrate the troubling degree to which long term success is associated with a student's demographic characteristics, regardless of the student's early academic prowess. Controlling for 3rd grade test achievement, poverty and race/ethnicity are strongly predictive of students' high school outcomes. Students who are EDS in the 3rd grade and in the 10th decile of the 3rdgrade achievement distribution score in high school math at the level of non-EDS students in the 9thdecile, are only about as likely to take an advanced math or science course in high school as non-EDS students in the 9th decile, and only about as likely to graduate as non-EDS students in the 2nd decile. In short, our models estimate the substantial magnitude of the academic headwinds that low-income students face over time.

We are careful not to imply that our findings are necessarily related directly or solely to students' experiences in schools themselves as there are disagreements about the degree to which schools ameliorate achievement gaps in different grades.[57] However, the combination of large achievement gaps in 3rd grade and the relationship between 3rd grade performance and long-term performance reinforces the challenge of reducing inequities in college readiness. Students in subgroups most likely to lag behind peers in 3rd grade tend to fall further behind over time rather than catching up.

It is certainly a judgment call as to whether the models we described here are *highly accurate* in predicting long term student outcomes, but there does appear to be broad agreement that tests ought to be used to diagnose when students are projected to struggle in their academic careers (NCLD, 2017; Richards et al., 2007). One might view schooling or other social service interventions as successful if they decrease the predictive power of 3rd grade tests, as this would imply that interventions are ensuring that early achievement does not become academic destiny. This suggests the need for more research along the lines of Austin et al. (2020) and Jang and Reardon (2019) that explore how students' educational trajectories may vary across different contexts, and, more importantly, why they may vary.

More novel is our exploration of using segments of achievement and parameters estimated from different states to predict high school outcomes. We find evidence that both using parameters generated from using segments of students' academic careers as well as using out-of-state generated parameters results in quite accurate estimates of students' high school outcomes. For instance, in the case of using segments, the correlation between .82-.99. And, while the accuracy of the estimates varies depending on the state pairings, the correlations between the estimates generated using own-state students to derive parameters and those generated using parameters derived from out-of-state students are also quite high: .78-.99 depending on state and outcome. These findings suggest that predictive modeling can be carried out successfully for more students, even in settings that lack the long panels of longitudinal data included in our analysis.

[57] See, for instance, recent evidence about the distribution of resources in schools across student subgroups (e.g., Goldhaber et al., 2018; Bischoff & Owens, 2019; Ijun Lai, 2020) and summer fall back and what it implies about differential student learning while students are in school (von Hipple et al., 2018).

The findings on different ways to make achievement projections have important implications for policy and practice. Of particular note is the use of these type of predictive models for state or district early warning systems, i.e., systems to highlight students who early on are in danger of not succeeding in high school. Our findings suggest that such systems likely need to target students for interventions far earlier than 8th grade as there is little that generally disrupts the trajectory that students are on when they are tested in the 3rd grade. But they also show that states that do not have data systems allowing them to estimate long-term educational outcomes (3rd grade to the end of high school) have good alternative options for generating predictions.

References

Abdiansah, A., & Wardoyo, R. (2015). Time complexity analysis of support vector machines (SVM) in LibSVM. *International journal computer and application*, 128(3), 28-34.

Allensworth, E. M., & Easton, J. Q. (2005). The on-track indicator as a predictor of high school graduation.

Allensworth, E. M., & Easton, J. Q. (2007). What Matters for Staying On-Track and Graduating in Chicago Public High Schools: A Close Look at Course Grades, Failures, and Attendance in the Freshman Year. Research Report. Consortium on Chicago School Research.

Austin, W., Figlio, D., Goldhaber, D., Hanushek, E., Kilbride, T., Koedel, C., Sean, J. L., Lou, J., Özek, U., Parsons, E., Rivkin, S., Sass, T., & Strunk, K. (2020). Where are initially low-performing students the most likely to succeed? A Multi-state Analysis of Academic Mobility (Preliminary Draft). *CALDER Working Paper* No. 227-0220.

Backes, B., Cowan, J., Goldhaber, D., Koedel, C., Miller, L. C., & Xu, Z. (2018). The common core conundrum: To what extent should we worry that changes to assessments will affect test-based measures of teacher performance? *Economics of Education Review*, 62, 48-65.

Betts, J. R., Hahn, Y., & Zau, A. C. (2017). Can testing improve student learning? An evaluation of the mathematics diagnostic testing project. *Journal of Urban Economics*, 100, 54-64.

Bontempi, G., Taieb, S. B., & Le Borgne, Y. A. (2012, July). Machine learning strategies for time series forecasting. *In European business intelligence summer school* (pp. 62-77). Springer, Berlin, Heidelberg.

Bischoff, K., & Owens, A. (2019). The segregation of opportunity: social and financial resources in the educational contexts of lower-and higher-income children, 1990–2014. Demography, 56, 1635-1664.

Bramer, M. (2007). Avoiding overfitting of decision trees. *Principles of data mining*, 119-134.

Burkam, D. T., & Lee, V. E. (2003). Mathematics, Foreign Language, and Science Course taking and the NELS: 88 Transcript Data. Working Paper No. 2003-01. *National Center for Education Statistics.*

Campaign, D. Q. (2009). The next step: Using longitudinal data systems to improve student success. Retrieved June, 19, 2021.

Campaign, D. Q. (2016). State progress. Retrieved March, 14, 2021.

Cawley, J., Heckman, J., & Vytlacil, E. (2001). Three observations on wages and measured cognitive ability. *Labour economics*, 8(4), 419-442.

Chajewski, M., Mattern, K. D., & Shaw, E. J. (2011). Examining the role of Advanced Placement® exam participation in 4-year college enrollment. *Educational Measurement: Issues and Practice*, 30, 16-27.

Castro, M., Expósito-Casas, E., López-Martín, E., Lizasoain, L., Navarro-Asencio, E., & Gaviria, J. L. (2015). Parental involvement on student academic achievement: A meta-analysis. *Educational research review*, 14, 33-46.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood (No. w17699). *National Bureau of Economic Research.*

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, 97, 31-47.

Cunha, F., Heckman, J. J., & Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3), 883-931.

Curtin, J., Hurwitch, B., & Olson, T. (2012). Development and Use of Early Warning Systems. SLDS Spotlight. *National Center for Education Statistics.*

De Hoyos, R., Ganimian, A. J., & Holland, P. A. (2017). Teaching with the test: experimental evidence on diagnostic feedback and capacity building for public schools in Argentina.

Desimone, L. (1999). Linking parent involvement with student achievement: Do race and income matter?. *The journal of educational research*, 93, 11-30.

Dougherty, C., Mellor, L., & Jian, S. (2006). The Relationship between Advanced Placement and College Graduation. 2005 AP Study Series, Report 1. *National Center for Educational Accountability.*

Dee, T. S. (2014). Stereotype threat and the student-athlete. Economic Inquiry, 52, 173-182.

Easton, J. Q., Johnson, E., & Sartain, L. (2017). The predictive power of ninth-grade GPA. *Chicago, IL: University of Chicago Consortium on School Research.*

Fiester, L. (2010). Early Warning! Why Reading by the End of Third Grade Matters. KIDS COUNT Special Report. *Annie E. Casey Foundation.*

Figlio, D., & Loeb, S. (2011). School accountability. *Handbook of the Economics of Education*, 3, 383-421.

Forte, D. (2021). *Why Parents, School Leaders, & Advocates Shouldn't Underestimate the Power of Statewide Assessments*. The Education Trust. https://edtrust.org/the-equity-line/why-parents-school-leaders-and-advocates-shouldnt-underestimate-the-power-of-statewide-assessments/

Geiser, S., & Santelices, M. V. (2007). Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes.

Goldhaber, D., Long, M. C., Person, A. E., Rooklyn, J., & Gratz, T. (2019). Sign Me Up: The Factors Predicting Students' Enrollment in an Early-Commitment Scholarship Program. *AERA Open*, 5, 2332858419857703.

Goldhaber, D., & Özek, U. (2019). How much should we rely on student test achievement as a measure of success?. Educational Researcher, 48(7), 479-483.

Goldhaber, D., Quince, V., & Theobald, R. (2018). How Did It Get This Way? Disentangling the Sources of Teacher Quality Gaps across Two States. Working Paper No. 209-1118-1. *National Center for Analysis of Longitudinal Data in Education Research (CALDER).*

Goldhaber, D., Theobald, R., & Fumia, D. (2018). Teacher Quality Gaps and Student Outcomes: Assessing the Association between Teacher Assignments and Student Math Test Scores and High School Course Taking. Working Paper 185. *National Center for Analysis of Longitudinal Data in Education Research (CALDER).*

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27, 83-85.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics*, 24, 411-482.

Henderson, A. T., & Berla, N. (1994). A new generation of evidence: The family is critical to student achievement.

Hernandez, D. J. (2011). Double Jeopardy: How Third-Grade Reading Skills and Poverty Influence High School Graduation. *Annie E. Casey Foundation.*

Ho, A. (2021). *A Smart Role for State Standardized Testing in 2021*. FutureEd. https://www.future-ed.org/a-smart-role-for-state-standardized-testing-in-2021/

Ijun Lai, W. Jesse Wood, Scott A. Imberman, Nathan Jones, Katharine Strunk (2020). Teacher Quality Gaps by Disability and Socioeconomic Status: Evidence from Los Angeles. *CALDER Working Paper* No. 228-0220

Jackson, J., & Cook, K. (2018). Modernizing California's Education Data System. *Public Policy Institute of California.*

Jang, H., & Reardon, S. F. (2019). States as sites of educational (in) equality: State contexts and the socioeconomic achievement gradient. Aera Open, 5(3), 2332858419872459.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13, 8-17.

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity?. *Educational researcher*, 31(1), 3-12.

Le Floch, K. C., Martinez, F., O'Day, J., Stecher, B., Taylor, J., & Cook, A. (2007). State and Local Implementation of the" No Child Left Behind Act." Volume III—Accountability under" NCLB" Interim Report. *US Department of Education*.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6), 321-332.

Massachusetts Department of Education (2018). Student Course Schedule (SCS) Data Handbook (Version 8.1 ed., p. 17). N.p.: Massachusetts DESE. Retrieved from http://www.doe.mass.edu/infoservices/data/scs/scs-datahandbook.docx

Mehana, M., & Reynolds, A. J. (1995). The Effects of School Mobility on Scholastic Achievement.

Mehana, M., & Reynolds, A. J. (2004). School mobility and achievement: A meta-analysis. *Children and Youth Services Review*, 26(1), 93-119.

Murnane, R. J., Willett, J. B., & Levy, F. (1995). The growing importance of cognitive skills in wage determination (No. w5076). *National Bureau of Economic Research.*

NCLD (2017, January 25). Identifying Struggling Students. In National Center for Learning Disabilities. Retrieved from https://www.ncld.org/research/state-of-learning-disabilities/identifying-struggling-students

NCES (2020, May). Public High School Graduation Rates. *National Center for Education Statistics*, May 2020, nces.ed.gov/programs/coe/indicator_coi.asp.

National Governors Association Center for Best Practices and Council of Chief State School Officers (NGAC and CCSSO). 2010. *Common core state standards*. Washington, DC: NGAC and CCSSO.

Neild, R. C., Balfanz, R., & Herzog, L. (2007). An early warning system. *Educational leadership*, 65, 28-33.

Pene, M. (2021b, March 15). *Standardized Testing Amid Pandemic Does Kids More Harm Than Good*. UT News. https://news.utexas.edu/2021/03/15/standardized-testing-amid-pandemic-does-kids-more-harm-than-good/

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, 79, 575-583.

Reardon, S. F. (2011). The widening socioeconomic status achievement gap: New evidence and possible explanations. *Whither opportunity*, 91-116.

Reardon, S. F. (2016). School segregation and racial academic achievement gaps. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 2(5), 34-57.

Richards, C., Pavri, S., Golez, F., Canges, R., & Murphy, J. (2007). Response to intervention: Building the capacity of teachers to serve students with learning difficulties. *Issues in Teacher Education*, 16, 55-64.

Roman, I., Santana, R., Mendiburu, A., & Lozano, J. A. (2020). In-depth analysis of SVM kernel learning and its components. *Neural Computing and Applications*, 1-20.

Rose, H., & Betts, J. (2004). The Effect of High School Courses on Earnings. *Review of Economics and Statistics*, 86(2), 497-513.

Sansone, D. (2019). Beyond early warning indicators: high school dropout and machine learning. *Oxford bulletin of economics and statistics*, 81(2), 456-485.

Silver, D., Saunders, M., & Zarate, E. (2008). What factors predict high school graduation in the Los Angeles Unified School District. Policy Brief, 14.

Sorensen, L. C. (2019). "Big data" in educational administration: An application for predicting school dropout risk. *Educational Administration Quarterly*, 55(3), 404-446.

Shores, K., & Steinberg, M. (2017). The impact of the great recession on student achievement: Evidence from population data. *Available at SSRN* 3026151.

Speroni, C. (2011). Determinants of Students' Success: The Role of Advanced Placement and Dual Enrollment Programs. An NCPR Working Paper. *National Center for Postsecondary Research.*

Strauss, V. (2015, March 1). The important things standardized tests don't measure. *The Washington Post*. https://www.washingtonpost.com/news/answer-sheet/wp/2015/03/01/the-important-things-standardized-tests-dont-measure/

Todd, P. E., & Wolpin, K. I. (2007). The production of cognitive achievement in children: Home, school, and racial test score gaps. Journal of Human capital, 1, 91-136.

Weaver-Randall, K., & Ireland, L. (2018). Graduation and Dropout Statistics Annual Report. Report to the Legislature [2016-17]. *Washington Office of Superintendent of Public Instruction.*

Wilder, S. (2014). Effects of parental involvement on academic achievement: A meta-synthesis. *Educational Review*, 66, 377-397.

Vanwinckelen, G., & Blockeel, H. (2012). On estimating model accuracy with repeated cross-validation. In *Benelearn 2012: Proceedings of the 21st belgian-dutch conference on machine learning*, 39-44.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7, 91.

von Hippel, P. T., Workman, J., & Downey, D. B. (2018). Inequality in reading and math skills forms mainly before kindergarten: A replication, and partial correction, of "Are Schools the Great Equalizer?". *Sociology of Education*, 91, 323-357.

Zau, A., & Betts, J. R. (2008). Predicting success, preventing failure: An investigation of the California high school exit exam. *Public Policy Instit. of CA.*

Tables and Figures

Table 1: Selected Descriptive Statistics on Analytic Sample.

Panel A: Characteristics by State and Quartile of Achivement

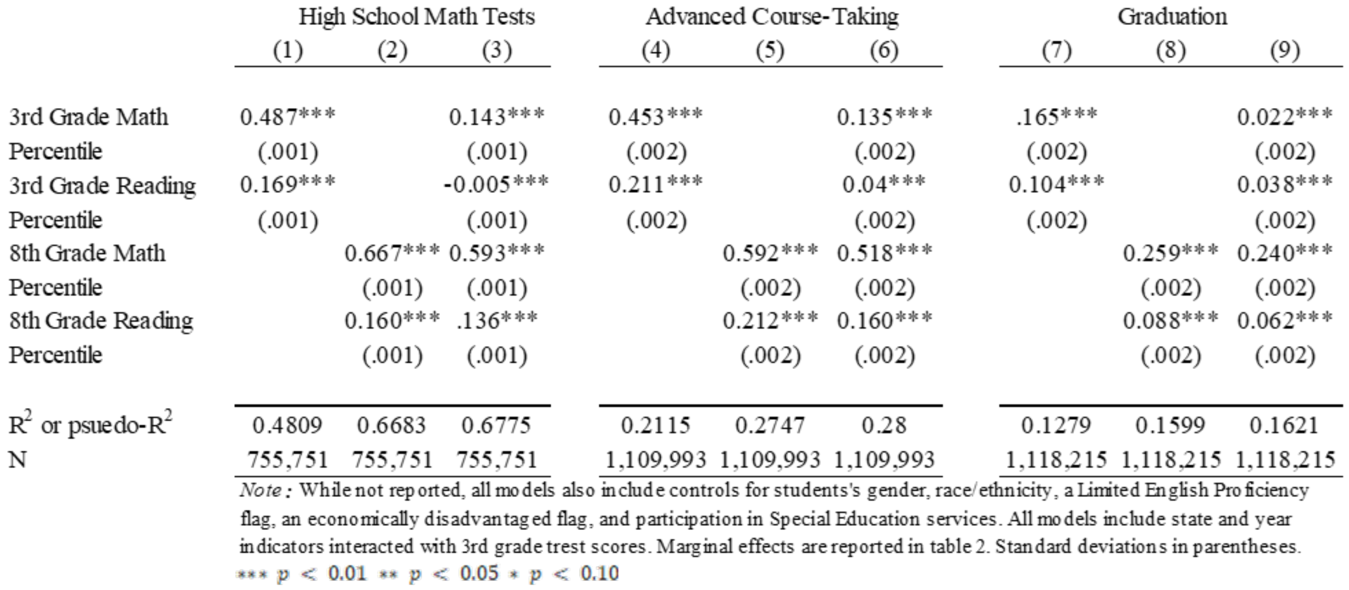
	State				Quartile of 3rd grade math achievement			
	Overall	MA	NC	WA	1 (Lowest)	2	3	4 (Highest)
Female	0.493	0.489	0.495	0.489	0.491	0.501	0.498	0.481
Amer. Ind.	0.014	0.003	0.015	0.022	0.021	0.016	0.012	0.007
Asian/PI	0.043	0.059	0.021	0.086	0.027	0.035	0.043	0.067
Black	0.186	0.078	0.273	0.051	0.323	0.212	0.138	0.068
Hispanic	0.122	0.155	0.087	0.185	0.193	0.137	0.098	0.057
White	0.602	0.679	0.573	0.614	0.403	0.567	0.676	0.769
LEP	0.075	0.089	0.060	0.104	0.141	0.081	0.049	0.028
EDS	0.435	0.352	0.450	0.468	0.662	0.495	0.363	0.212
SPED	0.105	0.173	0.073	0.133	0.225	0.100	0.058	0.033
Advanced	0.640	0.484	0.690	0.592	0.366	0.552	0.722	0.877
Graduated	0.860	0.914	0.849	0.857	0.726	0.847	0.902	0.947

Panel B: Total Size and Average Cohort Sizes by State

	Overall	MA	NC	WA
HS Math	755751	46647	77875	55266
Advanced	1109993	55248	66840	52252
Graduated	1118215	55341	67460	52532

Note: The Overall column includes students who are in the High School Math, Graduation, or Advanced Course sample. PI = Pacific Islander, LEP = Limited English Proficiency, EDS = Economically Disadvangtaged Student, SPED = Special Education

Table 2: Model Coefficients.



	High School Math Tests			Advanced Course-Taking			Graduation		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
3rd Grade Math	0.487***		0.143***	0.453***		0.135***	.165***		0.022***
Percentile	(.001)		(.001)	(.002)		(.002)	(.002)		(.002)
3rd Grade Reading	0.169***		-0.005***	0.211***		0.04***	0.104***		0.038***
Percentile	(.001)		(.001)	(.002)		(.002)	(.002)		(.002)
8th Grade Math		0.667***	0.593***		0.592***	0.518***		0.259***	0.240***
Percentile		(.001)	(.001)		(.002)	(.002)		(.002)	(.002)
8th Grade Reading		0.160***	.136***		0.212***	0.160***		0.088***	0.062***
Percentile		(.001)	(.001)		(.002)	(.002)		(.002)	(.002)
R^2 or psuedo-R^2	0.4809	0.6683	0.6775	0.2115	0.2747	0.28	0.1279	0.1599	0.1621
N	755,751	755,751	755,751	1,109,993	1,109,993	1,109,993	1,118,215	1,118,215	1,118,215

Note : While not reported, all models also include controls for students's gender, race/ethnicity, a Limited English Proficiency flag, an economically disadvantaged flag, and participation in Special Education services. All models include state and year indicators interacted with 3rd grade trest scores. Marginal effects are reported in table 2. Standard deviations in parentheses.

*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

Table 3: Model Coefficients by State (3rd grade).

	High School Math Tests			Advanced Course-Taking			Graduation		
	MA	NC	WA	MA	NC	WA	MA	NC	WA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
3rd Grade Math	0.479***	0.482***	0.540***	0.4972***	0.446***	0.449***	0.096***	0.191***	0.131***
Percentile	(0.002)	(0.002)	(0.004)	(0.005)	(0.002)	(0.005)	(0.004)	(0.002)	(0.004)
3rd Grade Reading	0.182	0.162***	0.172***	0.178***	0.222***	0.208***	0.055***	0.115***	0.105***
Percentile	(0.002)	(0.002)	(0.004)	(0.006)	(0.002)	(0.005)	(0.004)	(0.002)	(0.004)
R^2 or psuedo-R^2	0.533	0.451	0.578	0.150	0.238	0.149	0.138	0.131	0.098
N	233,236	467,249	55,266	165,743	735,241	209,009	166,023	742,063	210,129

Note : While not reported, all models also include controls for students's gender, race/ethnicity, a Limited English Proficiency flag, an economically disadvantaged flag, and participation in Special Education services. All models include year indicators interacted with 3rd grade trest scores. Marginal effects are reported in table 4. Standard deviations in parentheses.

*** $p < 0.01$ ** $p < 0.05$ * $p < 0.10$

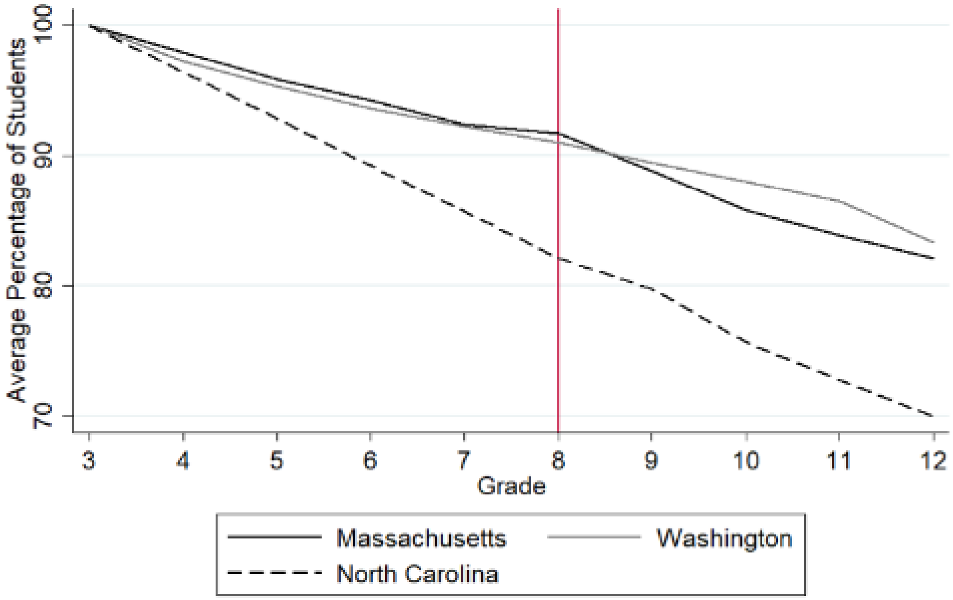
Table 4: Correlations of Predicted High School Math Percentile by Grade and State.

		Overall			MA			NC			WA		
		3rd	8th	3rd & 8th	3rd	8th	3rd & 8th	3rd	8th	3rd & 8th	3rd	8th	3rd & 8th
Overall	3rd	1											
	8th	0.7925	1										
	3rd & 8th	0.8462	0.9936	1									
MA	3rd	0.9974	0.7918	0.8439	1								
	8th	0.7941	0.9988	0.9928	0.7954	1							
	3rd & 8th	0.8473	0.9929	0.999	0.8472	0.9939	1						
NC	3rd	0.9974	0.7892	0.8443	0.9938	0.7898	0.8448	1					
	8th	0.7902	0.9992	0.9932	0.7884	0.9977	0.9918	0.7889	1				
	3rd & 8th	0.8438	0.9928	0.9994	0.8405	0.9914	0.9979	0.8441	0.9936	1			
WA	3rd	0.9962	0.7901	0.843	0.9958	0.794	0.8462	0.9938	0.7873	0.8405	1		
	8th	0.7967	0.9986	0.9926	0.7973	0.9994	0.9933	0.7924	0.9976	0.9913	0.7977	1	
	3rd & 8th	0.8509	0.9925	0.9986	0.8503	0.9934	0.9994	0.8485	0.9917	0.9977	0.8512	0.9939	1

Table 5: Correlations between Model Predictions by State.

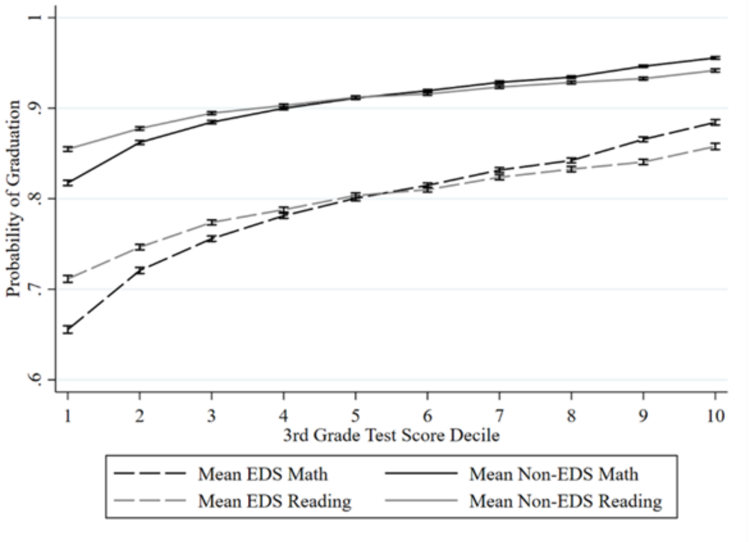
		MA data			NC data			WA data		
		MA model	NC model	WA model	MA model	NC model	WA model	MA model	NC model	WA model
Graduation	MA model	1			1			1		
	NC model	0.849	1		0.929	1		0.862	1	
	WA model	0.938	0.891	1	0.958	0.835	1	0.915	0.892	1
Advanced course-	MA model	1			1			1		
	NC model	0.870	1		0.954	1		0.893	1	
	WA model	0.971	0.857	1	0.939	0.936	1	0.966	0.887	1
HS math	MA model	1			1			1		
	NC model	0.993	1		0.995	1		0.992	1	
	WA model	0.995	0.993	1	0.997	0.995	1	0.993	0.991	1

Figure 1: Average Percent Student Sample Attrition by Grade and State, 2006-2018.



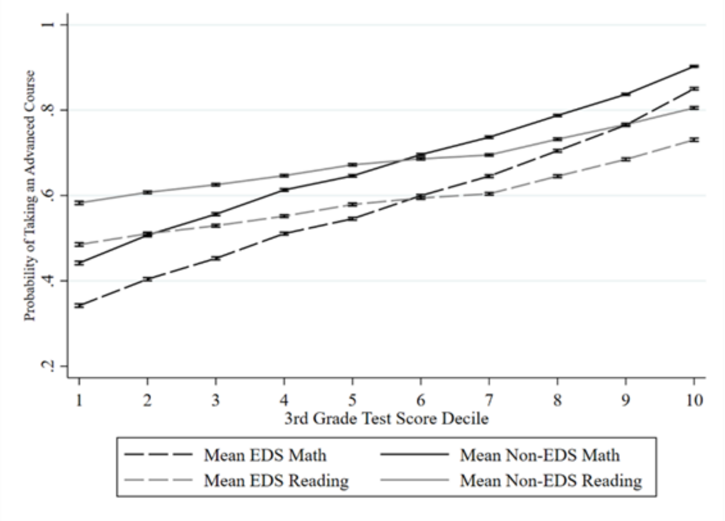
Note: Average percent of observable 3rd grade students throughout the K-12 education system, broken up by state. Most student's observable in eighth grade have the *Above 50* outcome, and those observable through 12th grade have the *Graduation* outcome.

Figure 2: Probability of Graduation by 3rd grade Test Score Decile and EDS.



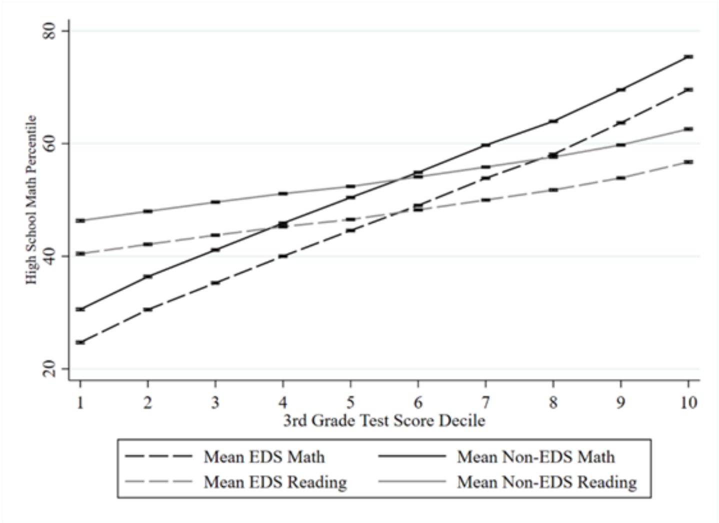
Note: Probability of graduation by 3rd grade test score decile and EDS, estimated as marginal effects. Consistent, large effects of EDS status are seen, lowering probability of graduation by eight percent to 10% across all test scores for math and reading—approximately the same effect as going from the first to tenth decile of scores.

Figure 3: Probability of Advanced Course-Taking by 3rd grade Test Score Decile and EDS.



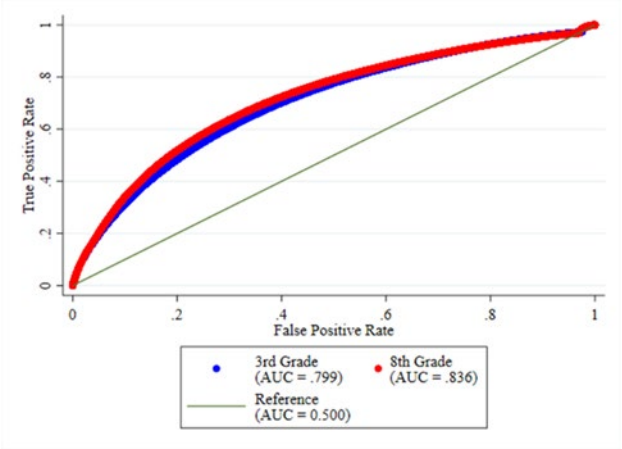
Note: Probability of advanced course-taking by 3rd grade test score decile and EDS, estimated as marginal effects. Relatively consistent, large effects of EDS status are seen, lowering probability of advanced course-taking by eight percent to 10% across all test scores for math and reading—approximately the same effect as improving test scores by one decile.

Figure 4: High School Math Percentile by 3rd grade Test Scores and EDS.



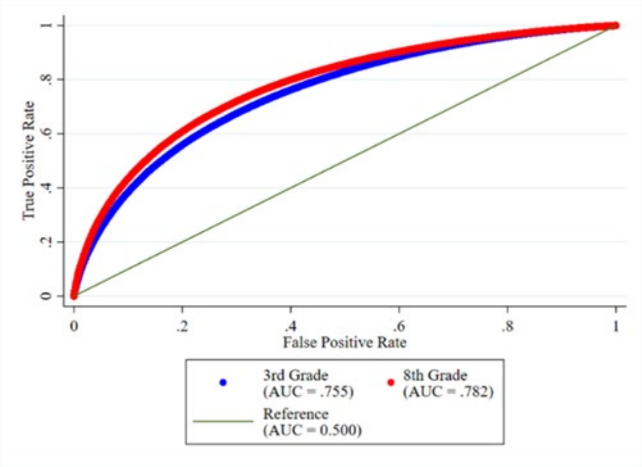
Note: High school math percentile by 3rd grade test score decile and EDS, estimated as marginal effects. Relatively consistent, large effects of EDS status are seen, lowering high school math percentile by three percent to five percent across all test scores for math and reading—a slightly smaller effect as improving test scores by one decile

Figure 5: ROC Curve Predicting Graduation using Third and Eighth Grade Test Scores.



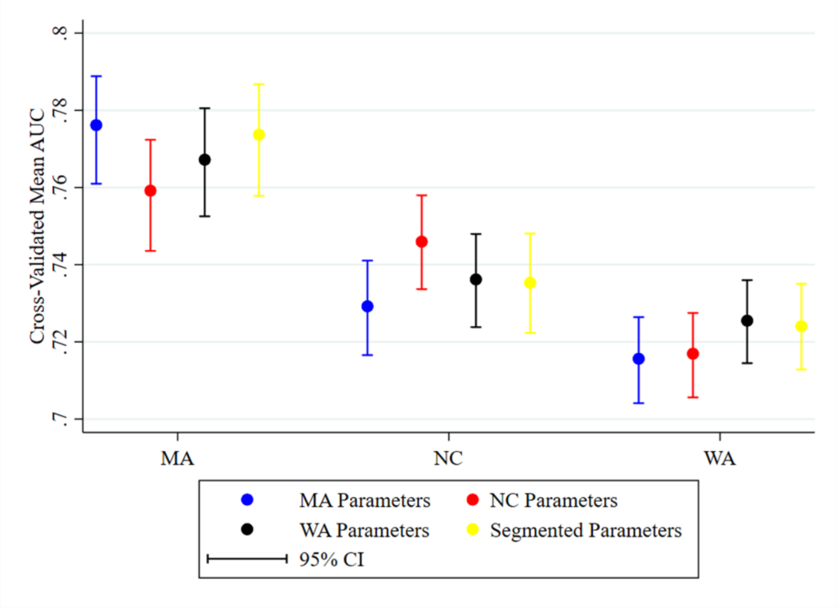
Note: ROC curves corresponding to graduation prediction using both 3rd grade test scores and eighth grade test scores, with reported AUCs in the legend. The similarity of AUC and general shape of ROC curve shows a strong capacity for effective intervention targeting early in students' academic careers—as early as 3rd grade.

Figure 6: ROC Curve Predicting Advanced Course-Taking using Third and Eight Grade Test. Scores



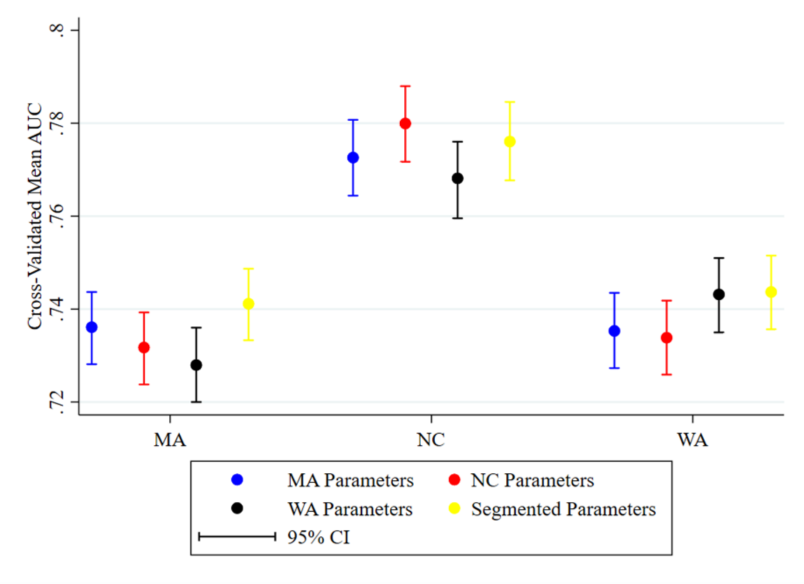
Note: ROC curves corresponding to advanced course-taking prediction using both 3^{rd} grade test scores and eighth grade test scores, with reported AUCs in the legend. The similarity of AUC and general shape of ROC curve shows a strong capacity for predicting high achievement as early as 3rd grade.

Figure 7: Graduation Cross-Validated AUC Estimates by Prediction Model.



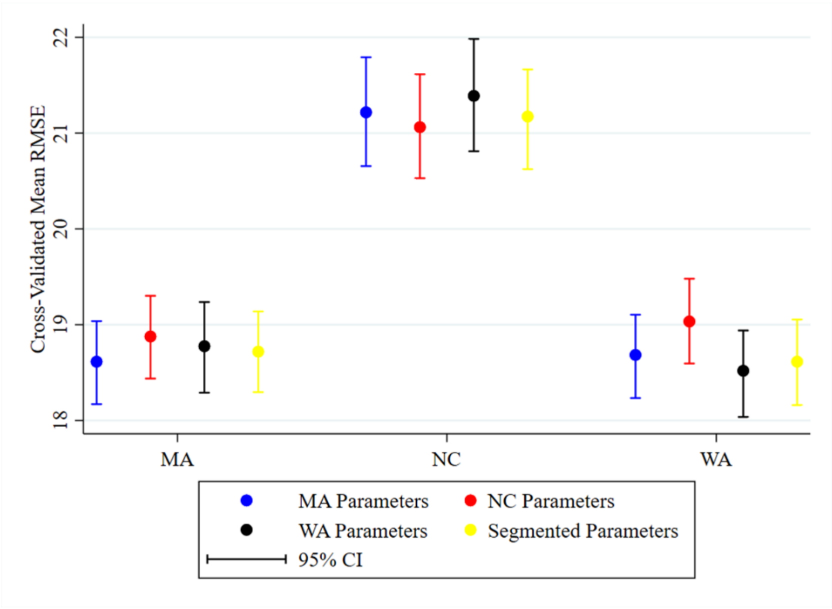
Note: Mean estimated probabilities of 10-fold cross-validated AUC for graduation. Confidence intervals are generated by repeating 10-fold CV over 100 iterations.

Figure 8: Advanced Course-Taking Cross-Validated AUC Estimates by Prediction Model.



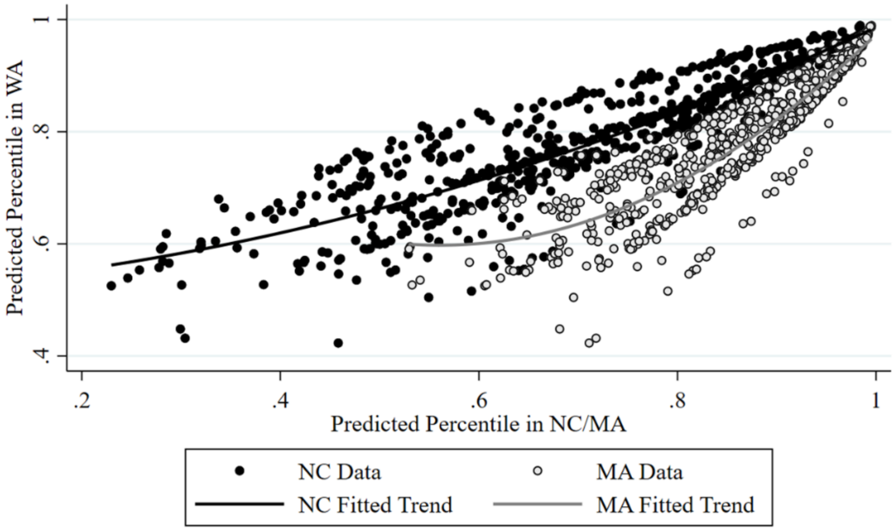
Note: Mean estimated probabilities of 10-fold cross-validated AUC for advanced course-taking. Confidence intervals are generated by repeating 10-fold CV over 100 iterations.

Figure 9: High School Math Percentile Cross-Validated RMSE Estimates by Prediction Model.



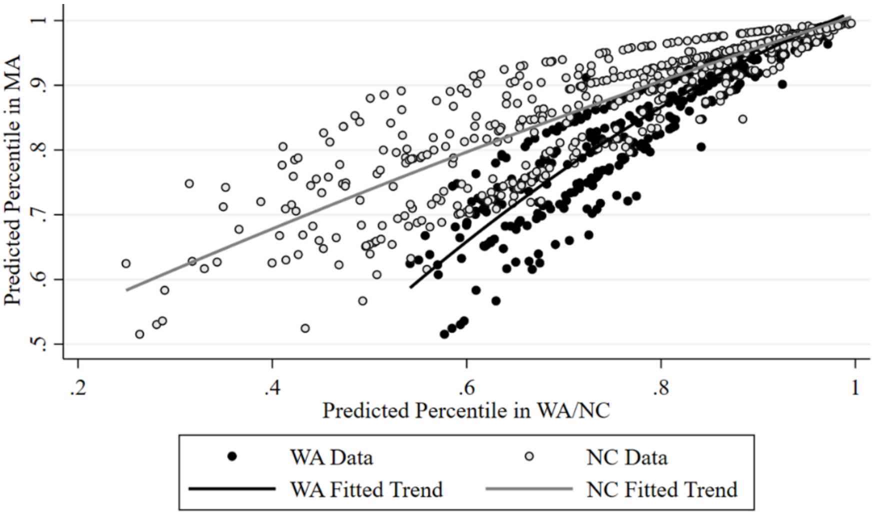
Note: Mean estimates of 10-fold cross-validated RMSE for high school math tests. Confidence intervals are generated by repeating 10-fold CV over 100 iterations.

Figure 10: Scatterplot of Predicted Probabilities of Graduation in WA vs Predicted Probabilities from Out-of-State Models (3rd grade).



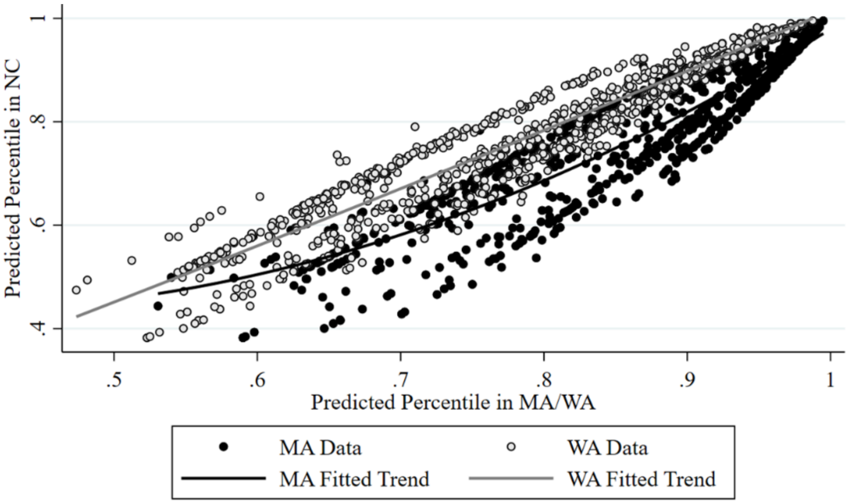
Note: Scatterplot of predicted probabilities of graduation in Washington compared to predicted probabilities in North Carolina and Massachusetts, estimated on students in Washington State. Points displayed are a random subset of less than five percent of the data, where the probability of displaying a point is inversely proportional to the predicted probability of graduation for readability.

Figure 11: Scatterplot of Predicted Probabilities of Graduation in MA vs Predicted Probabilities from Out-of-State Models (3rd grade).



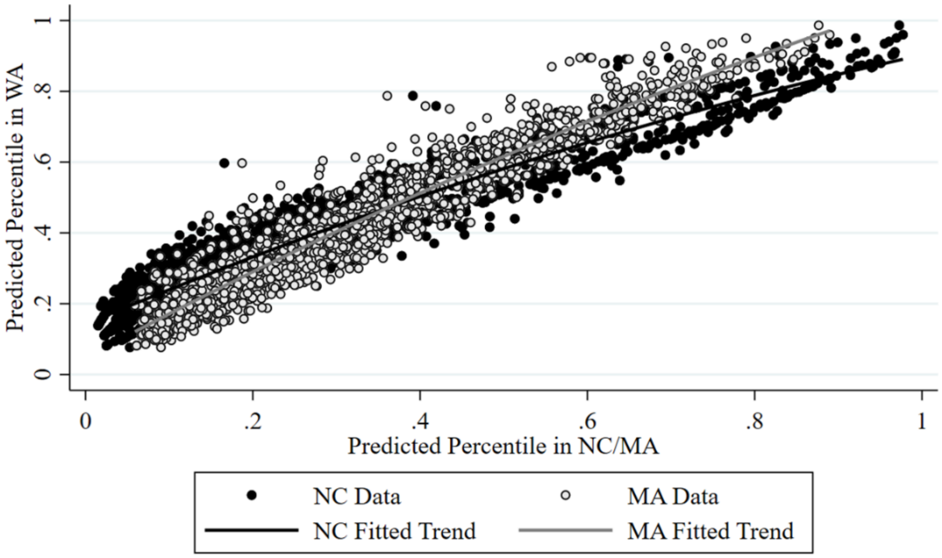
Note: Scatterplot of predicted probabilities of graduation in Massachusetts compared to predicted probabilities in North Carolina and Washington, estimated on students in Massachusetts. Points displayed are a random subset of less than five percent of the data, where the probability of displaying a point is inversely proportional to the predicted probability of graduation for readability.

Figure 12: Scatterplot of Predicted Probabilities of Graduation in NC vs Predicted Probabilities from Out-of-State Models (3rd grade).



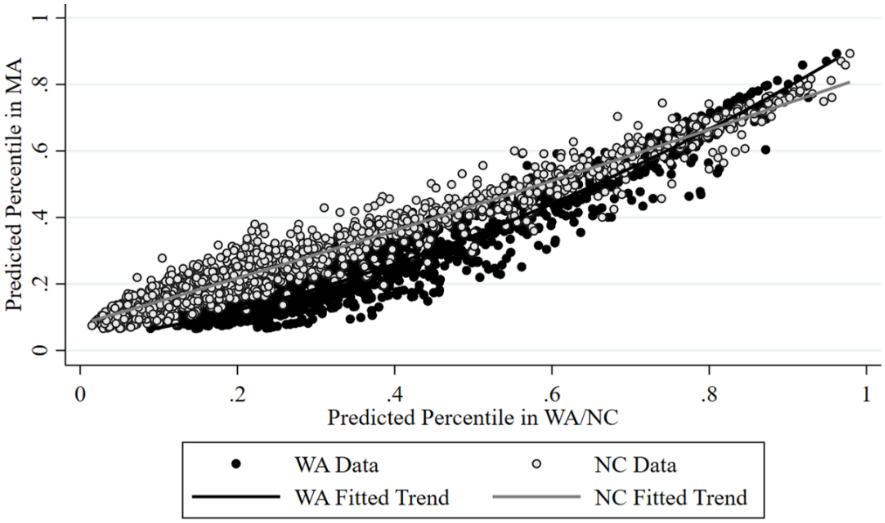
Note: Scatterplot of predicted probabilities of graduation in North Carolina compared to predicted probabilities in Massachusetts and Washington, estimated on students in North Carolina. Points displayed are a random subset of less than five percent of the data, where the probability of displaying a point is inversely proportional to the predicted probability of graduation for readability.

Figure 13: Scatterplot of Predicted Probabilities of Advanced Course-Taking in WA vs Predicted Probabilities from Out-of-State Models (3rd grade).



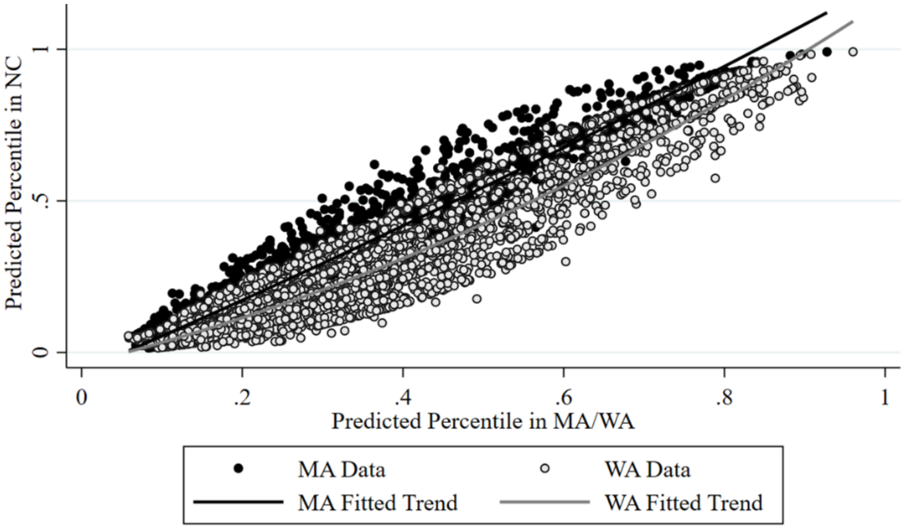
Note: Scatterplot of predicted probabilities of advanced course-taking in Washington compared to predicted probabilities in Massachusetts and North Carolina, estimated on students in Washington. Points displayed are a random subset of less than five percent of the data, where the probability of displaying a point is inversely proportional to the predicted probability of advanced course-taking for readability.

Figure 14: Scatterplot of Predicted Probabilities of Advanced Course-Taking in MA vs Predicted Probabilities from Out-of-State Models (3rd grade).



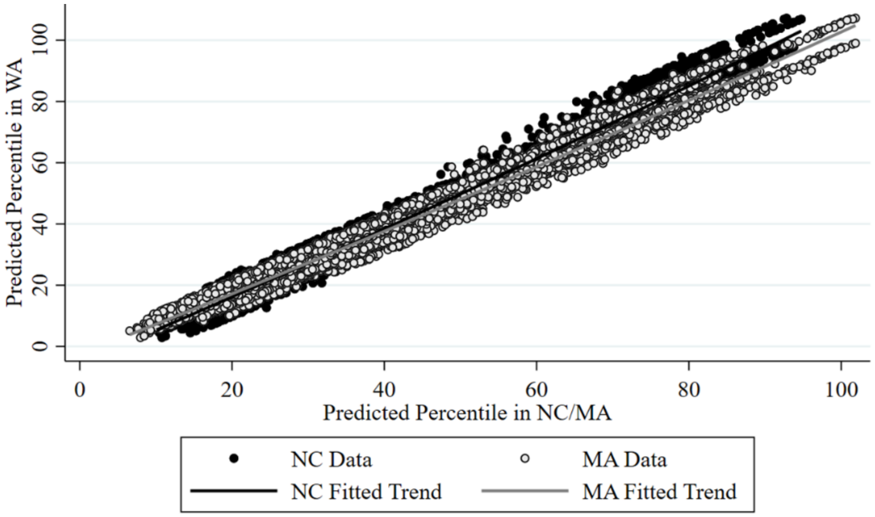
Note: Scatterplot of predicted probabilities of advanced course-taking in Massachusetts compared to predicted probabilities in Washington and North Carolina, estimated on students in Massachusetts. Points displayed are a random subset of less than five percent of the data, where the probability of displaying a point is inversely proportional to the predicted probability of advanced course-taking for readability.

Figure 15: Scatterplot of Predicted Probabilities of Advanced Course-Taking in NC vs Predicted Probabilities from Out-of-State Models (3rd grade).



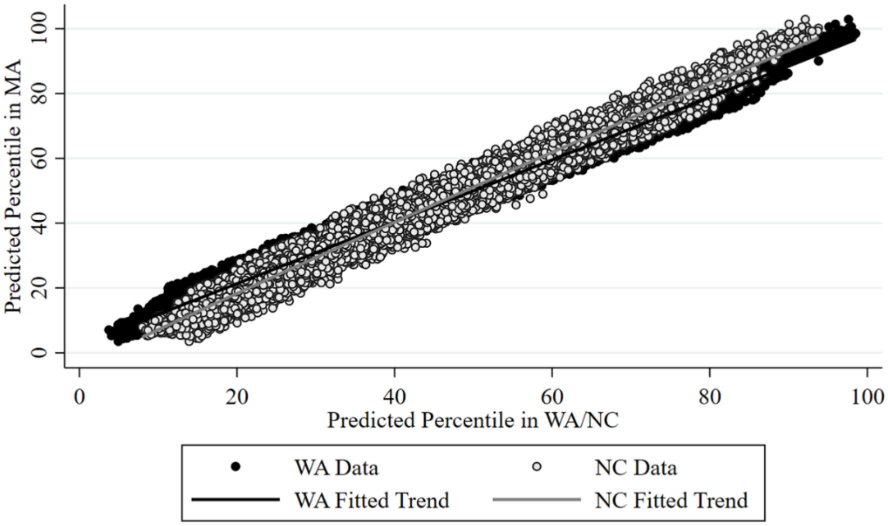
Note: Scatterplot of predicted probabilities of advanced course-taking in North Carolina compared to predicted probabilities in Washington and Massachusetts, estimated on students in North Carolina. Points displayed are a random subset of less than five percent of the data, where the probability of displaying a point is inversely proportional to the predicted probability of advanced course-taking for readability.

Figure 16: Scatterplot of Predicted High School Math Percentile in WA vs Predicted Probabilities from Out-of-State Models (3rd grade).



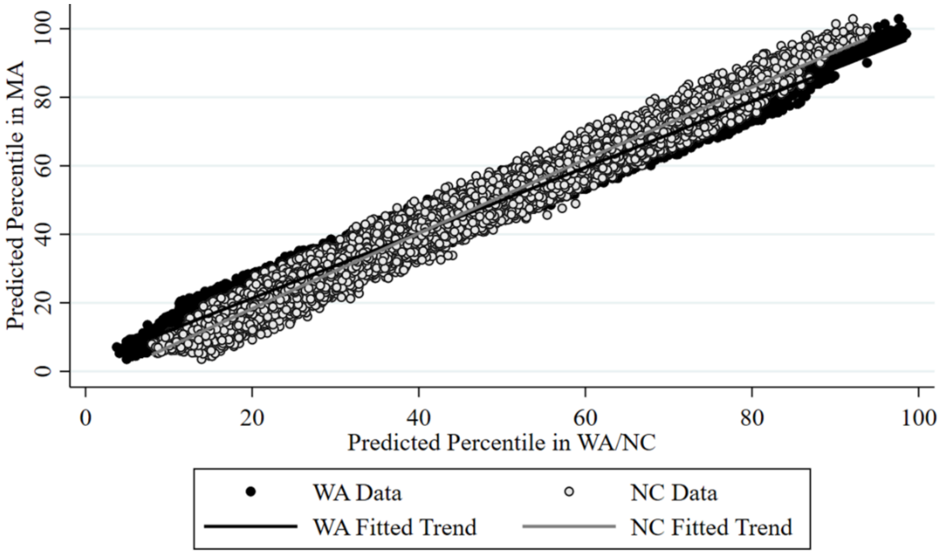
Note: Scatterplot of predicted percentiles of high school math test in Washington compared to predicted probabilities in North Carolina and Massachusetts, estimated on students in Washington. Points displayed are a random subset of less than five percent of the data.

Figure 17: Scatterplot of Predicted High School Math Percentile in MA vs Predicted Probabilities from Out-of-State Models (3rd grade).



Note: Scatterplot of predicted percentiles of high school math test in Massachusetts compared to predicted probabilities in North Carolina and Washington, estimated on students in Massachusetts. Points displayed are a random subset of less than five percent of the data.

Figure 18: Scatterplot of Predicted High School Math Percentile in NC vs Predicted Probabilities from Out-of-State Models (3rd grade).



Note: Scatterplot of predicted percentiles of high school math test in North Carolina compared to predicted probabilities in Massachusetts and Washington, estimated on students in North Carolina. Points displayed are a random subset of less than five percent of the data.

Appendix A

A.1 The Effect of Nonlinear Classifiers on Predictive Capacity

We use three ML approaches for predicting high school outcomes: Kernel Support Vector Machines, Random Forest Classification, and Gradient Boosted Decision Trees. We describe how we use these three approaches in estimating the schooling outcomes below. Without loss of generality, we describe the method in terms of students' graduation as the outcome.

Support Vector Machines

Given N students $(\boldsymbol{X}_i, Y_i)_{i\in 1,..,n}$ where $Y_i \in \{-1, 1\}$ represents is a dichotomization of graduation, Support Vector Machines (SVMs) estimate a hyperplane which maximizes the distance between graduate and non-graduate students according to their observable characteristics. This can be written as two parallel hyperplanes

$$X\beta - \beta_0 \geq \ 1, \qquad Y_i = \ \ 1 \qquad \text{(A1a)}$$
$$X\beta - \beta_0 \leq -1, \qquad Y_i = -1 \qquad \text{(A1b)}$$

Since the distance between these two hyperplanes is $\frac{2}{||\beta||}$, maximizing the distance between is equivalent to minimizing $||\beta||$. This is equivalent to solving the optimization problem

$$\min||\beta|| \quad s.t. \qquad Y_i(X_i\beta + \beta_0) \geq 1, \quad \text{for all } 1 \leq i \leq N. \qquad \text{(A2)}$$

The optimal values for β and β_0 are completely determined by students with the closest characteristics X_i that end up with different graduation outcomes. In other words, the resulting boundary between graduation and non-graduation is completely determined by students with observable characteristic values most ambiguously related to their graduation.

While we omit the details here, this procedure can be generalized to non-linearly separable data using "kernel functions" by describing more general distances between points (e.g., Roman et al., 2020). Since kernel support vector machines are based entirely on the distance between student characteristics, and not the characteristics themselves, they tend to handle high dimensional data very well. Moreover, because of their low-dimensional parameterization, kernel support vector machines avoid overfitting more than other methods. However, these models do have some drawbacks. They mainly perform well when there is clear separation between graduates and non-graduates according to student observables, and their accuracy depends largely on the arbitrary choice of an appropriate kernel, model estimation has a computational complexity that's cubic in the number of students N, and they do not produce any interpretable probabilities of membership.

Classification and Regression Trees, Random Forests, and Gradient Boosting

Classification and Regression Tree methods operate by iteratively splitting students along a single dimension of their observable characteristics according to what best separates graduates from non-graduates. This iterative splitting results in each student falling into a "bin" based on their observable characteristics, with the end goal that the majority of students in the same bin having the same graduation outcome. In the context of binary outcomes, such as graduation, the agreement associated with a tree can be measured with "Gini Impurity", a weighted average of picking the wrong class:

$$I_G(P) = 2p(1-p) \tag{A3}$$

for $p = P(\text{Graduation})$. The Gini Impurity tends to zero as the grouped data is better described by a single outcome and is maximized when outcomes are uniformly distributed among data in the group ($p \to 1/2$). Hence, this will be minimized when the majority of students in each "bin" are either graduates or non-graduates.

To make this idea concrete, we first introduce the KL-Divergence of two probability distributions P and Q for a binary outcome:

$$KL(P\,||Q) = p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right), \tag{A4}$$

a non-negative measure of agreement which is 0 if and only if $P = Q$ almost everywhere. The Gini impurity can be minimized by minimizing the negative KL-divergence between P and $\frac{1}{2}$, which is equal to minimizing

$$H(Y) := -KL(P||\tfrac{1}{2}) = -p\log(p) - (1-p)\log(1-p), \tag{A5}$$

known in the literature as the entropy of the graduation outcome Y with distribution P. Finally, we define the expected "information gain" conditioned on student characteristics X as

$$E_X[IG(Y,X)] = E_X[H(Y) - H(Y|X)] = H(Y) - E_X[H(Y|X)], \tag{A6}$$

where $H(Y|X)$ is the entropy of the student graduation outcome Y using conditional probabilities of the form $P(\text{Graduated} \mid\mid X)$. A Classification and Regression Tree for student graduation would then be optimized by constructing a model for $P(\text{Graduated} \mid\mid X)$ that iteratively finds variable and cut-point pairs which maximize the expected information gain in equation (C4).

As might be obvious, the solution to iterative optimization of (C4) can lead to overfitting. For example, if the data are continuous, one could construct a tree so specific as to classify each student separately. A technique called Bootstrap Aggregating (bagging) attempts to alleviate this issue by selecting a random sample with replacement of the training set and fits trees to the samples. Specifically, for some number of bootstrap repetitions B:

For $b = 1, \ldots, B$:

1. Sample, with replacement, N students
2. Train classification or regression tree $f_b(\boldsymbol{X}, Y)$ on the subsampled students

Predictions are then made by averaging over each of the estimates $\widehat{f_b}$. For an out of sample student i with characteristics X_i, the prediction is taken to be the median value of $\{\widehat{f_b}(\boldsymbol{x}'), b \in B\}$. Though a single estimated classification tree can be highly sensitive to training data, the aggregation of many such models is not as long as the models are sufficiently uncorrelated. This "Bagging" procedure helps to both decorrelate these models and reduce overfitting issues. To further reduce the correlation across predictions $\widehat{f_b}$, a technique called "Random Forests" conduct bootstrap aggregating of classification trees with each sample of students $(X_i, Y_i)_{i \in b}$

only containing a random subset of their characteristics X (typically $\sqrt{p}$ for $X \in \mathbb{R}^p$). The reduction in overfitting allows "Random Forest" methods to retain high prediction accuracy of classification trees out-of-sample by alleviating overfitting.

An alternative method for boosting classification tree accuracy is called Gradient Boosting. At the m^{th} step, gradient boosting methods fit a classification tree $f_m(\boldsymbol{x})$ by improving the information gain (C4) based on residuals from the classification $f_{m-1}(\boldsymbol{x})$. This update has the form

$$f_m(X) = f_{m-1}(X) + \nu\, f_m{}'(X), \qquad \text{(A7)}$$

where $f_m'(X)$ is an "adjustment" model based on the residuals between the true graduation outcome Y and the estimated graduation outcome from $f_{m-1}(X)$, and $\nu < 0.1$ is a "learning rate" parameter to reduce overfitting.

Table A1: AUCs by Prediction Model

	Graduation			Advanced Course-Taking		
Model	(1)	(2)	(3)	(4)	(5)	(6)
Probit	0.75	0.76	0.76	0.77	0.80	0.82
Interaction	0.76	0.78	0.79	0.80	0.83	0.83
SVM	0.74	0.76	0.75	0.82	0.84	0.85
Random Forest	0.81	0.83	0.84	0.83	0.85	0.87
Gradient Boost	0.79	0.82	0.83	0.83	0.85	0.87
3rd Grade Tests	X		X	X		X
8th Grade Tests		X	X		X	X
N	1,120,023	1,120,023	1,120,023	1,110,873	1,110,873	1,110,873

Notes: AUC Measures for Graduation and Advanced Course-Taking for various prediction models for all three states. State and year effects are omitted for the purpose of valid out-of-sample prediction.

A.3 The Confounding Effects of Schoolyear

As students' academic environment change over time, the relationship between their observable characteristics out outcomes may also change, and any prediction model will be unable to capture non-linear time trends. Due to the long panel of data in North Carolina, ranging from 1998 to 2012 for 3rd graders able to be identified as advanced course-takers or high school graduates,[58] we can explore potential changes in the magnitude of the relationship between students' early academic testing distribution and high school outcomes over a long period. We model the relationship between 3rd grade test scores and dichotomous high school high school outcomes for students in the initial cohorts of the sample (1998-2003) and the final cohorts of the sample (2008-2012) using equation (2) and compare the coefficients. Then, we estimate equation (2) on a random subset of the early cohorts (1998-2003) and use the model to predict on out-of-sample students in the early cohorts as well as future cohorts (2008-2012).

[58] Due to the only recently available high school test score data in North Carolina, we omit this outcome from the robustness check.

Table A1 shows coefficients estimating the relationship between math and reading 3rd grade test score percentile and dichotomous high school outcomes. While statistically significant, we find that the magnitude of the relationship between math percentile score and probability of high school outcome stays relatively consistent over time. However, the relationship between 3rd grade reading score in early and future cohorts in North Carolina does change substantively.

Table A2: Math and Reading Percentile Coefficients by Year Range

	Advanced Course-Taking		Graduation	
	(1)	(2)	(3)	(4)
	'98-'03	'08-'12	'98-'03	'08-'12
3rd Grade Math	1.59***	1.86***	1.01***	0.96***
	(0.01)	(0.03)	(0.01)	(0.04)
3rd Grade Reading	0.84***	1.14***	0.75***	0.33***
	(0.01)	(0.03)	(0.01)	(0.04)
Student Controls	X	X	X	X
N	447,614	87,034	408,149	76,455

Notes: All regressions control for student observables. Standard deviations in parentheses.
$*** \ p < 0.01 \ ** \ p < 0.05 \ * \ p < 0.10$

When predicting future cohorts' outcomes using coefficients estimated from early cohorts, we find that there is very little decrease in overall prediction accuracy. The AUC of predictive future cohort graduation with within-year range student data is .764 compared to an AUC of .726 when using early cohort relationships with graduation. Similarly, the AUC of within-year range student data for advanced course-taking is .795, compared to an AUC of .829 when using early cohort relationships with advanced course-taking. This suggests the relationships hold for multiple years.

A.4 The Effect of Additional Grade-Specific Test Scores on Predictive Capacity

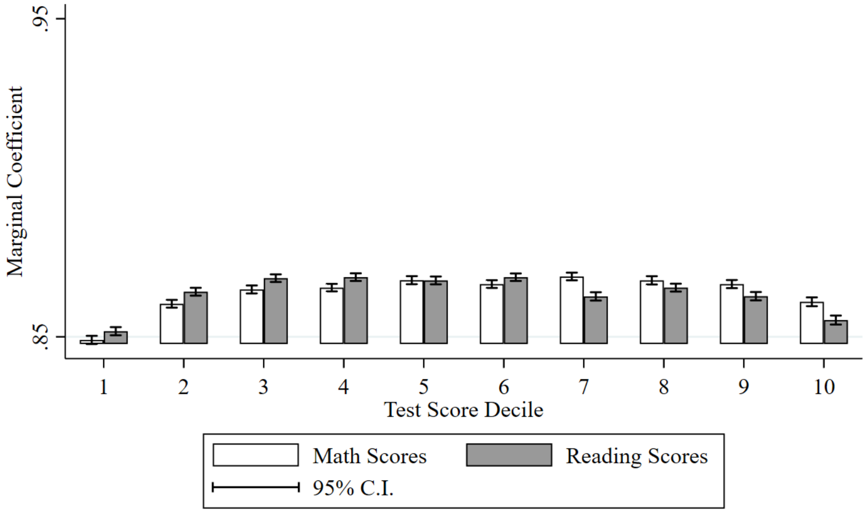
Table A3: Predictive Accuracy Measures by Test Score Combination

	(1) HS Math Tests	(2) Advanced Course-Taking	(3) Graduation
Grade	RMSE	AUC	
3rd	20.2	0.77	0.74
4th	19.5	0.78	0.74
5th	18.8	0.79	0.75
3rd & 4th	19.1	0.78	0.75
3rd & 5th	18.5	0.79	0.75
4th & 5th	18.4	0.79	0.75
3rd, 4th & 5th	18.3	0.79	0.75
N	728,014	1,075,108	1,082,708

Notes: Root mean squared error and AUC measures by test score combination. State and year effects are omitted for the purpose of valid out-of-sample prediction.

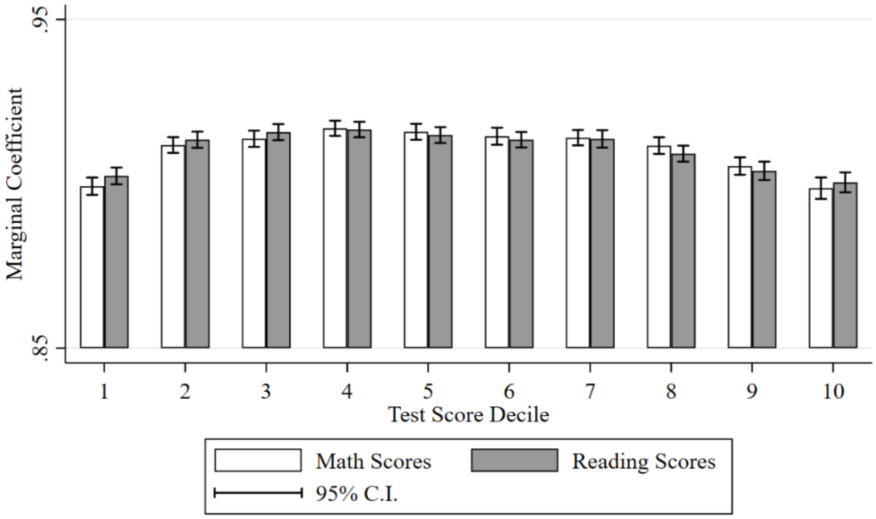
A.5 Sample Persistence

Figure A1: Marginal Effects of Test Score on Sample Persistence by 8th Grade



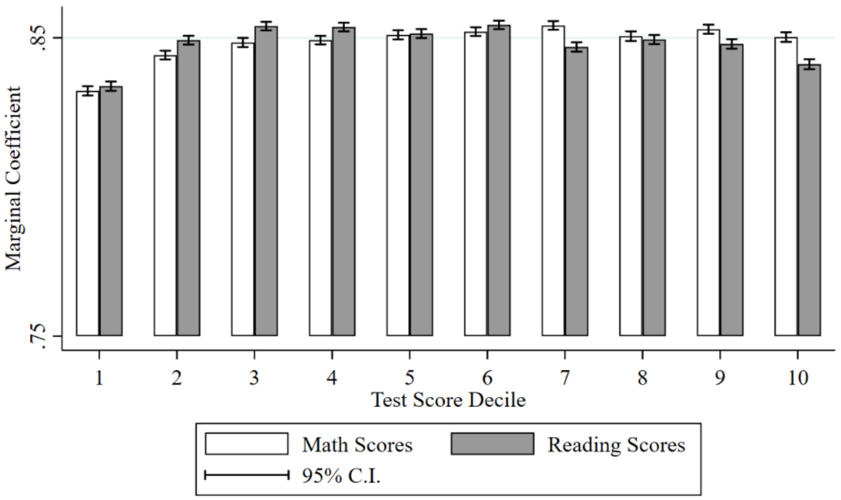
Notes: Marginal probabilities of sample persistence by 8th grade, controlling for student effects, broken up by test score decile. Students in the lowest decile of test scores are significantly less likely to persist in the sample through 8th grade, and students in thehighest decile of test scores are somewhat less likely to persist in the sample through 8th grade.

Figure A2: Marginal Effects of Test Score on Sample Persistence by 8th Grade (MA)



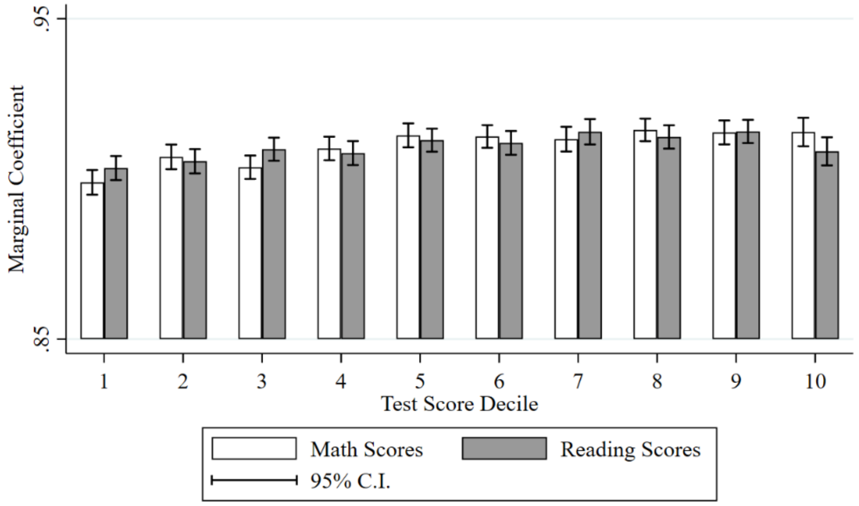
Notes: Marginal probabilities of sample persistence by 8th grade in MA, controlling for student effects, broken up by test score decile. Students in the lowest and highest decile of test scores are significantly less likely to persist in the sample through 8th grade.

Figure A3: Marginal Effects of Test Score on Sample Persistence by 8th Grade (NC)



Notes: Marginal probabilities of sample persistence by 8th grade in NC, controlling for student effects, broken up by test score decile. Students in the lowest decile of test scores are significantly less likely to persist in the sample through 8th grade.

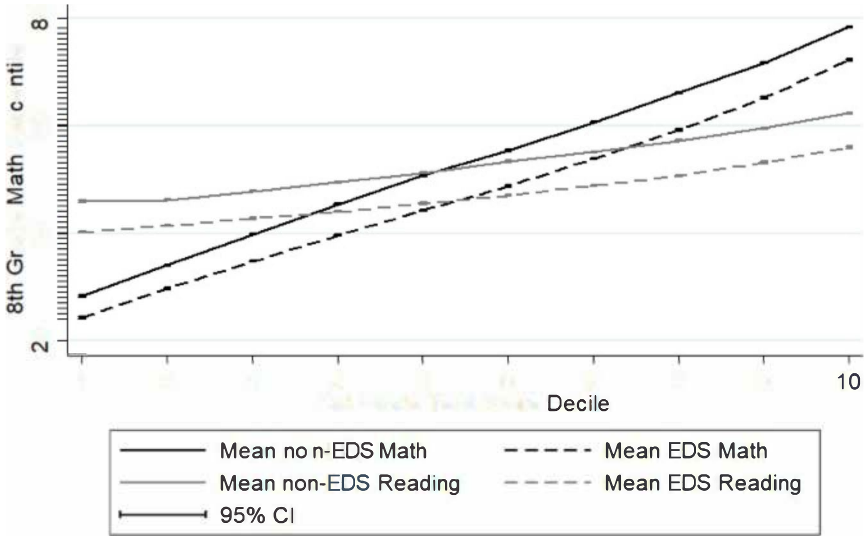
Figure A4: Marginal Effects of Test Score on Sample Persistence by 8th Grade (WA)



Notes: Marginal probabilities of sample persistence by 8th grade in WA, controlling for student effects, broken up by test score decile. Probability of sample persistence is relatively consistent across test score decile.

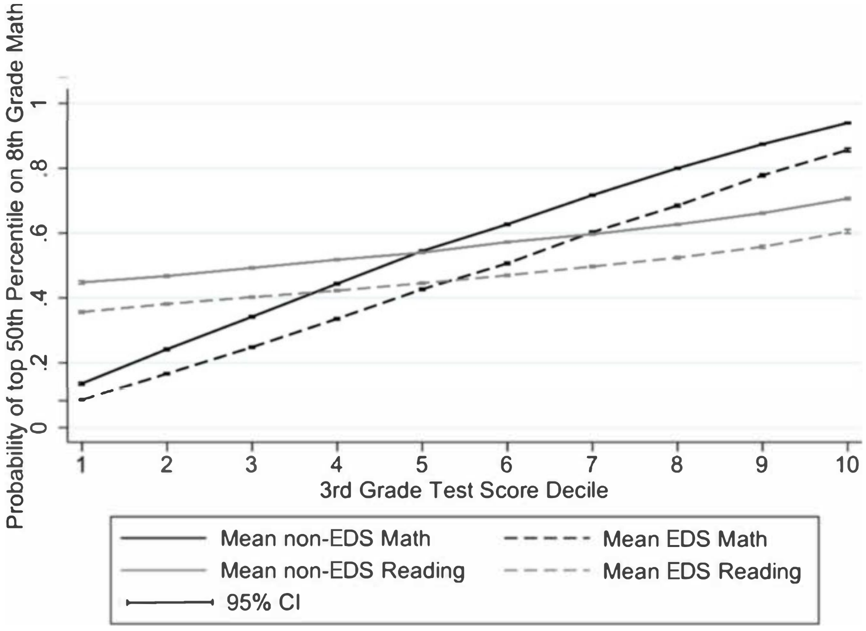
A.6 Supplementary Outcome Results

Figure A4: 8th Grade Math Percentile by 3rd Grade Test Scores and EDS



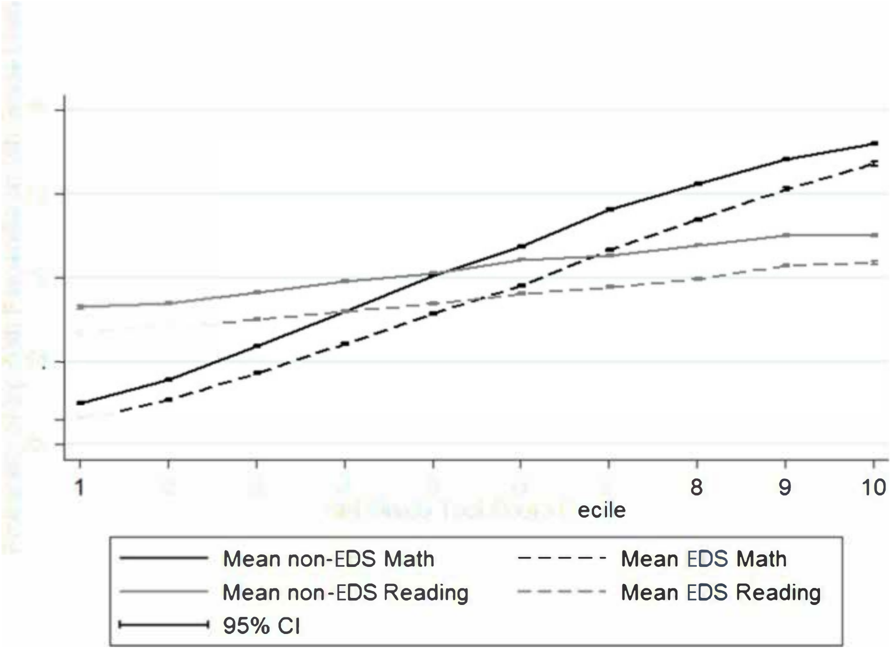
Notes: 8th grade math percentile by 3rd grade test score decile and EDS, estimated as marginal effects. Large effects of EDS status are seen, lowering percentile by up to 10 for math and reading.

Figure A5: Probability of Top 50th Percentile in 8th Grade Math by 3rd Grade Test Scores and EDS



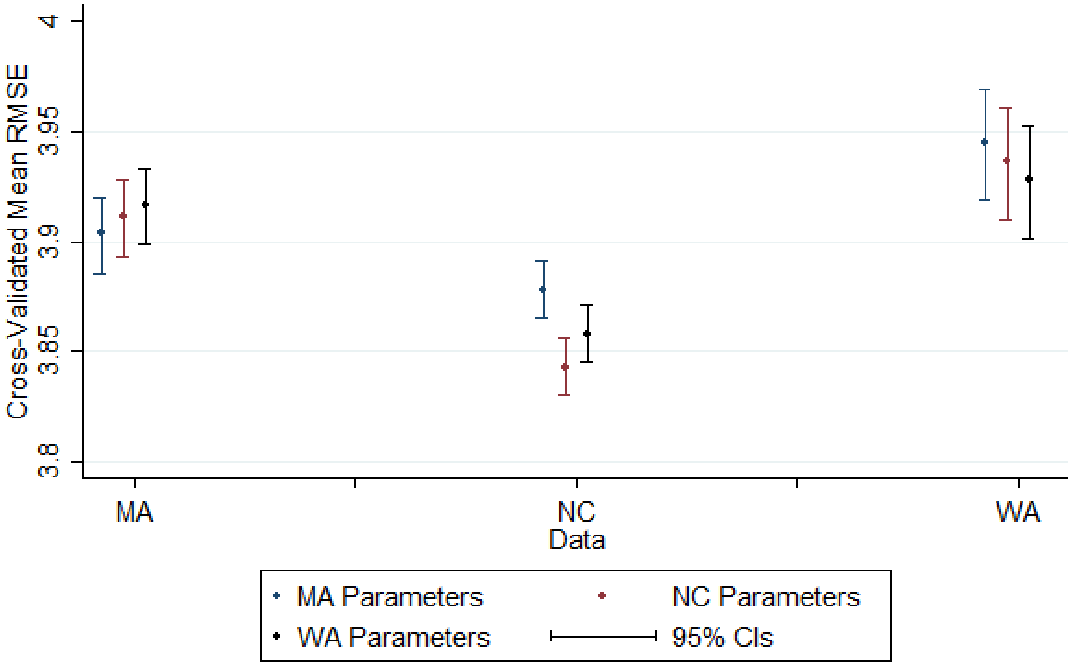
Notes: Probability of top 50th percentile in 8th grade math by 3rd grade test score decile and EDS, estimated as marginal effects. Large effects of EDS status are seen, lowering probability of top-half achievement by up to 10% for math and reading—approximately the same effect as a one-decile change in math test score.

Figure A6: Probability of Top 50th Percentile in High School Math by 3rd Grade Test Scores and EDS



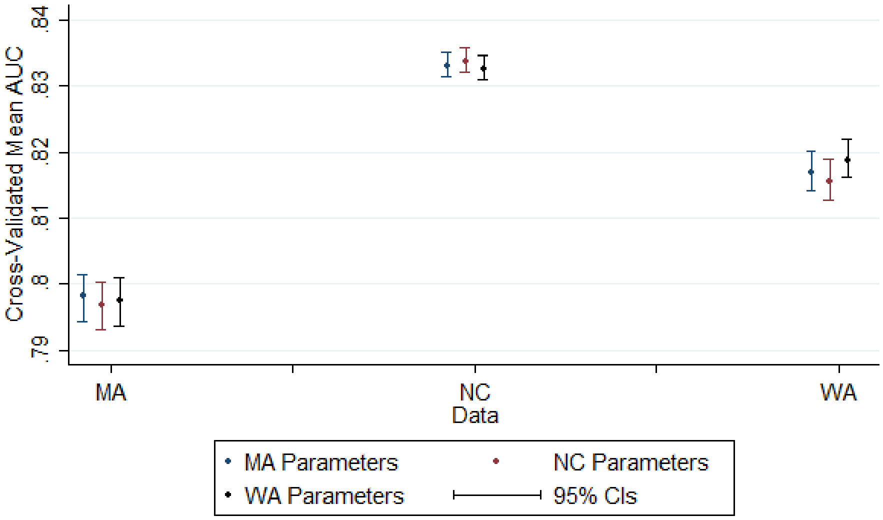
Notes: 8th grade math percentile by 3rd grade test score decile and EDS, estimated as marginal effects. Large effects of EDS status are seen, lowering percentile by up to 10 for math and reading.

Figure A7: 8th Grade Math Percentile Cross-Validated RMSE Estimates by Prediction Model



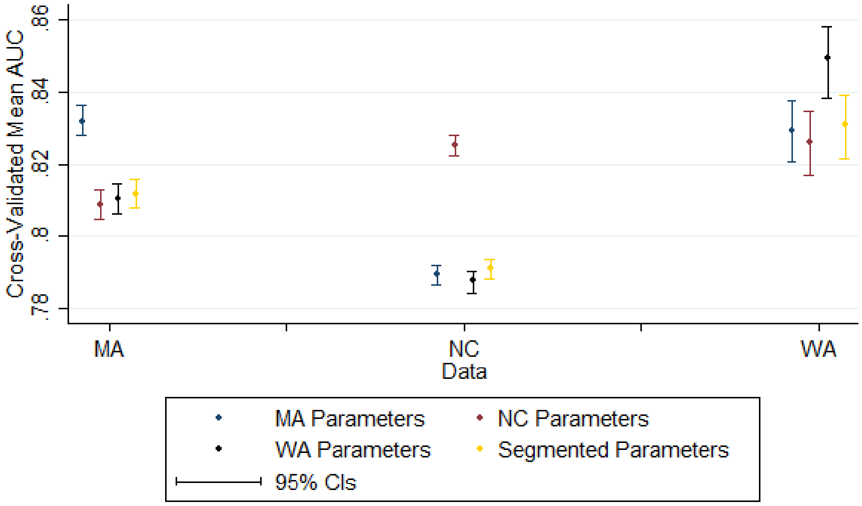
Notes: Mean estimates of 10-fold cross-validated RNSE for 8th grade math tests. Confidence intervals are generated repeating 10-fold CV over 100 iterations.

Figure A8: Probability of Top 50th Percentile in 8th Grade Math Cross-Validated AUCEstimated by Prediction Model



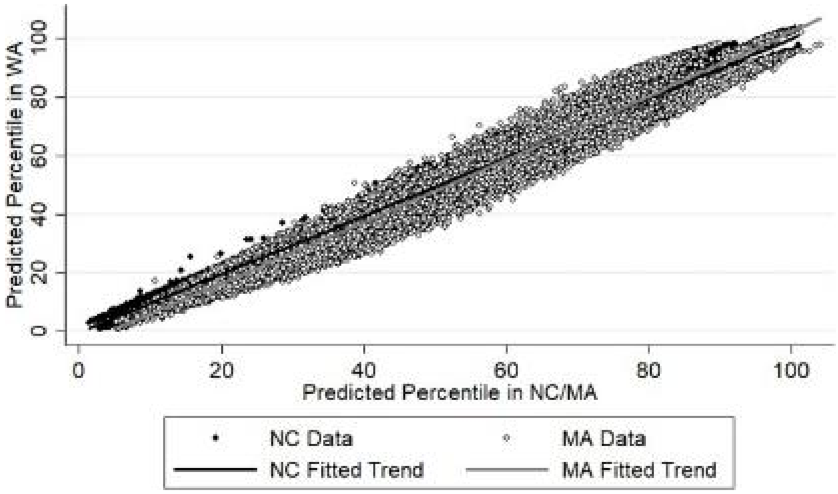
Notes: Mean estimates of 10-fold cross-validated AUC for the probability of scoring in the top half of 8th grade math test scores. Confidence intervals are generated by repeating 10-fold CV over 100 iterations

Figure A9: Probability of Top 50th Percentile High School Math Cross-Validated AUCEstimated by Prediction Model



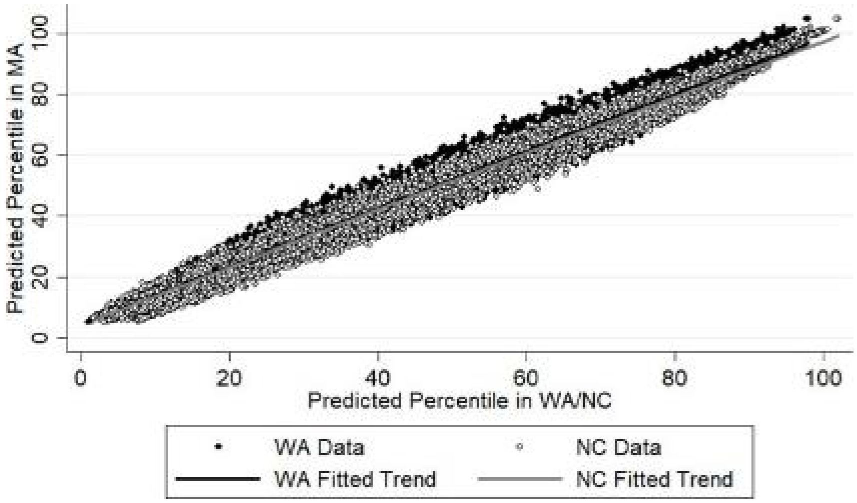
Notes: Mean estimates of 10-fold cross-validated AUC for the probability of scoring in the top half of high school math test scores. Confidence intervals are generated by repeating 10-fold CV over 100 iterations.

Figure A10: Scatterplot of Predicted 8th Grade Math Percentile in WA vs Predicted Probabilitiesfrom Out-of-State Models (3rd Grade)



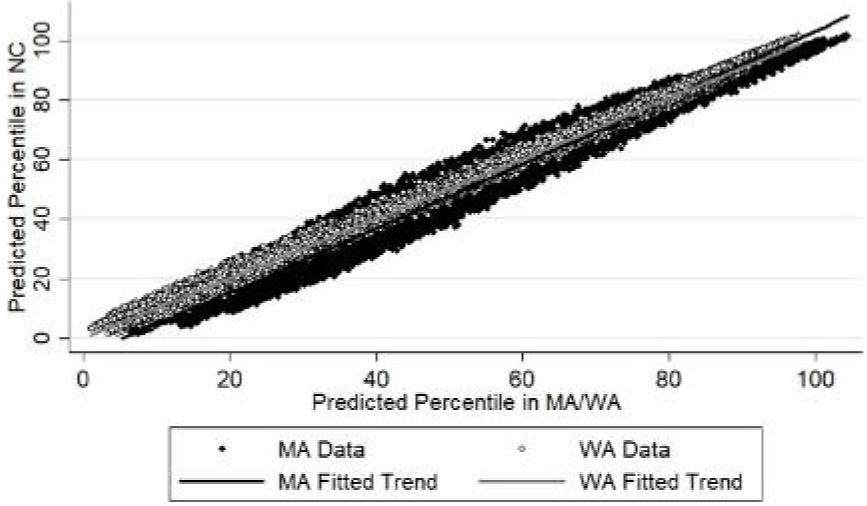
Notes: Scatterplot of predicted percentiles of 8th grade math test in Washington compared to predicted probabilities in North Carolina and Massachusetts, estimated on students in Washington.

Figure A11: Scatterplot of Predicted 8th Grade Math Percentile in MA vs Predicted Probabilitiesfrom Out-of-State Models (3rd Grade)



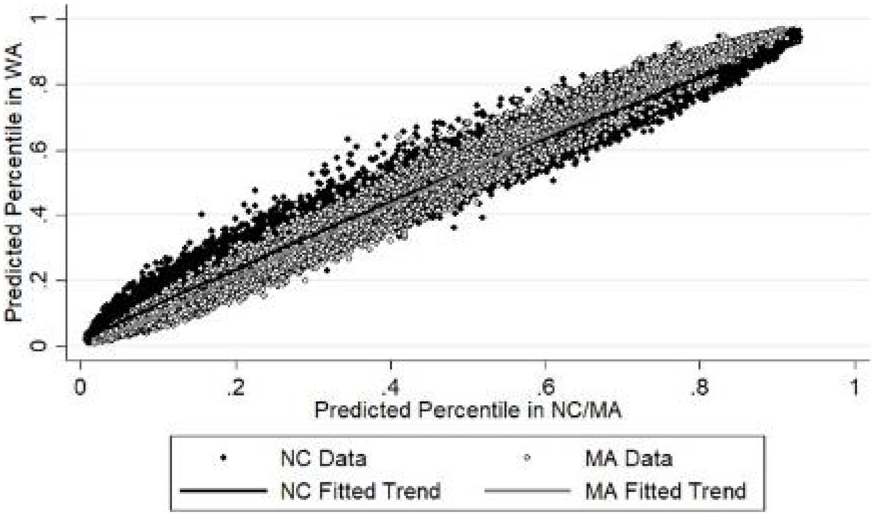
Notes: Scatterplot of predicted percentiles of 8th grade math test in Massachusetts compared to predicted probabilities in North Carolina and Washington, estimated on students in Massachusetts.

Figure A12: Scatterplot of Predicted 8th Grade Math Percentile in NC vs Predicted Probabilitiesfrom Out-of-State Models (3rd Grade)



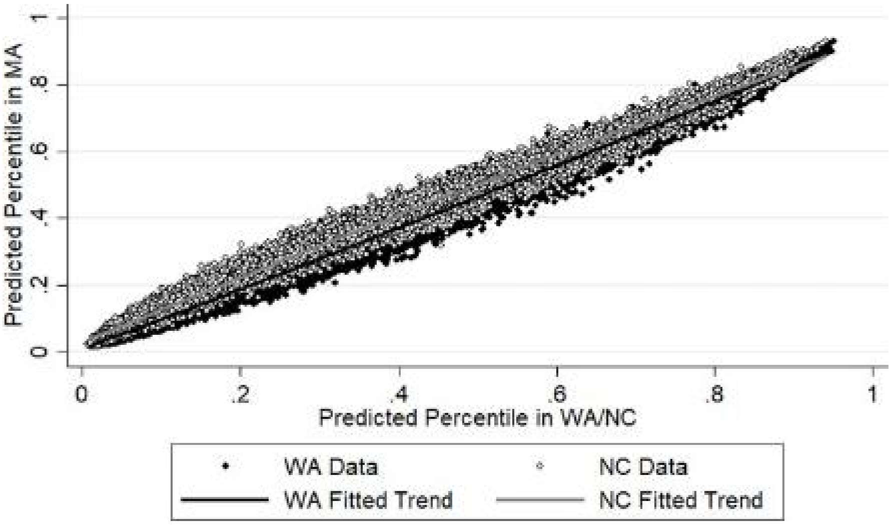
Notes: Scatterplot of predicted percentiles of 8th grade math test in North Carolina compared to predicted probabilities in Massachusetts and Washington, estimated on students in North Carolina.

Figure A13: Scatterplot of Predicted Probabilities of Top Half 8th Grade Math Tests in WA vsPredicted Probabilities from Out-of-State Models (3rd Grade)



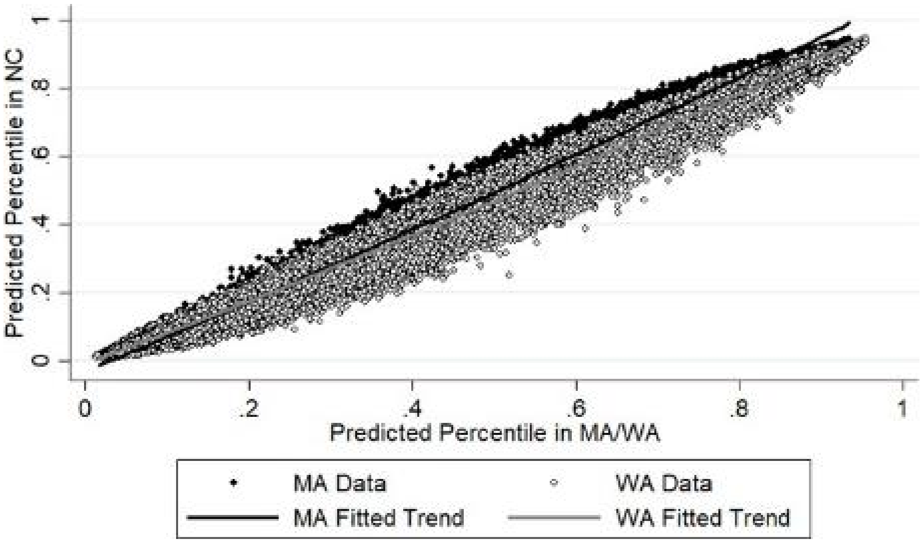
Notes: Scatterplot of predicted probability of scoring in the top half of 8th grade math tests in Washington compared to predicted probabilities in North Carolina and Massachusetts, estimated on students in Washington.

Figure A14: Scatterplot of Predicted Probabilities of Top Half 8th Grade Math Tests in MA vsPredicted Probabilities from Out-of-State Models (3rd Grade)



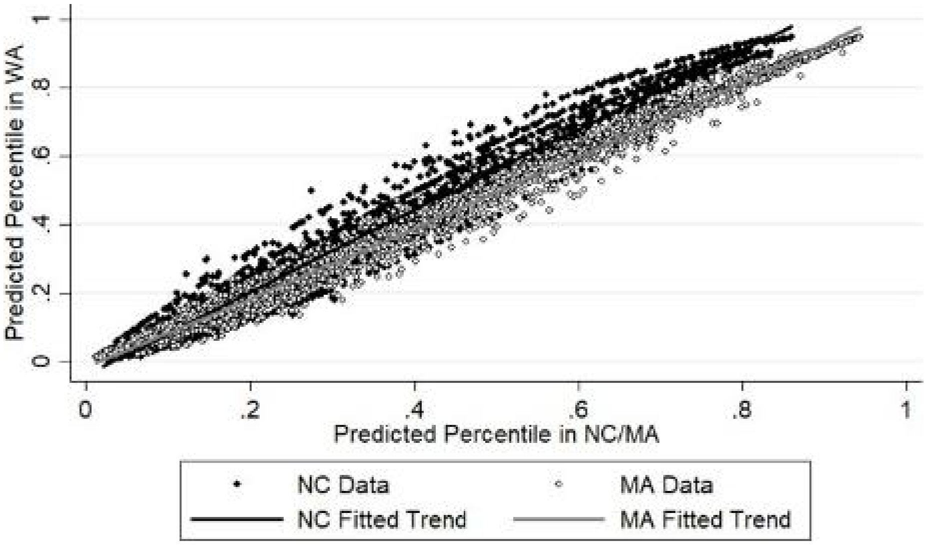
Notes: Scatterplot of predicted probability of scoring in the top half of 8th grade math tests in Massachusetts compared to predicted probabilities in North Carolina and Washington, estimated on students in Massachusetts.

Figure A15: Scatterplot of Predicted Probabilities of Top Half 8th Grade Math Tests in NC vsPredicted Probabilities from Out-of-State Models (3rd Grade)



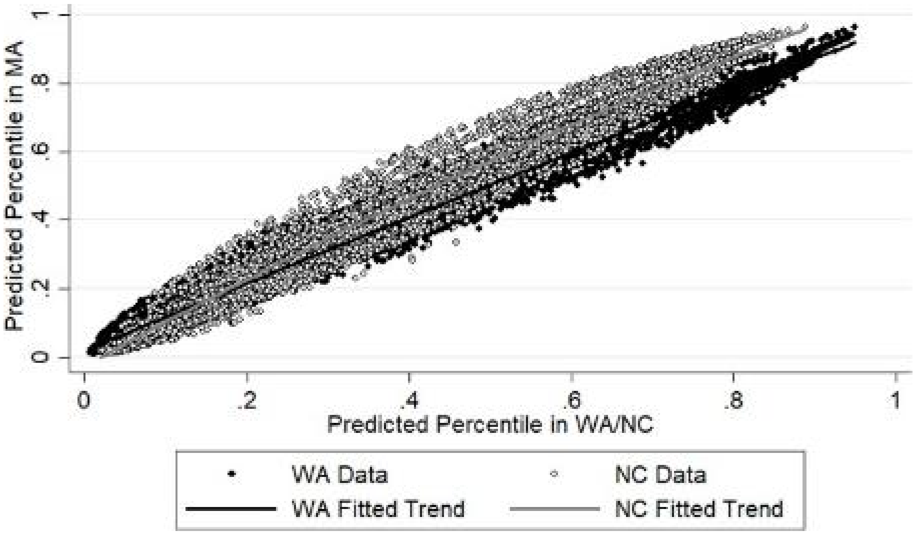
Notes Scatterplot of predicted probability of scoring in the top half of 8th grade math tests in North Carolina compared to predicted probabilities in Massachusetts and Washington, estimated on students in North Carolina.

Figure A16: Scatterplot of Predicted Probabilities of Top Half High School Math Tests in WA vs Predicted Probabilities from Out-of-State Models (3rd Grade)



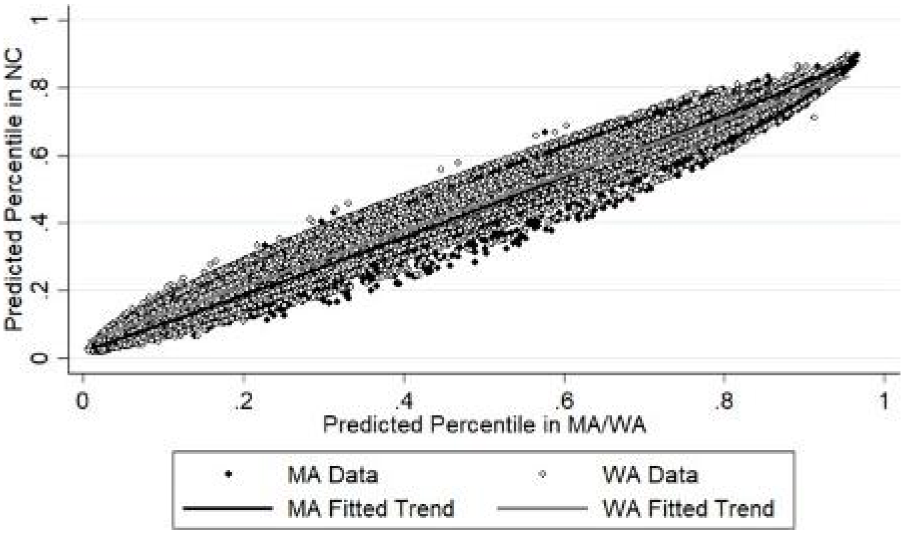
Notes: Scatterplot of predicted probability of scoring in the top half of high school math tests in Washington compared to predicted probabilities in North Carolina and Massachusetts, estimated on students in Washington.

Figure A17: Scatterplot of Predicted Probabilities of Top Half High School Math Tests in MA vs Predicted Probabilities from Out-of-State Models (3rd Grade)



Notes: Scatterplot of predicted probability of scoring in the top half of high school math tests in Massachusetts compared to predicted probabilities in Washington and North Carolina, estimated on students in Massachusetts.

Figure A18: Scatterplot of Predicted Probabilities of Top Half High School Math Tests in NC vs Predicted Probabilities from Out-of-State Models (3rd Grade)



Notes: Scatterplot of predicted probability of scoring in the top half of high school math tests in North Carolina.

A.6 State-Specific Results

Figure A19: ROC Curve of Graduation by Model Specification in Massachusetts

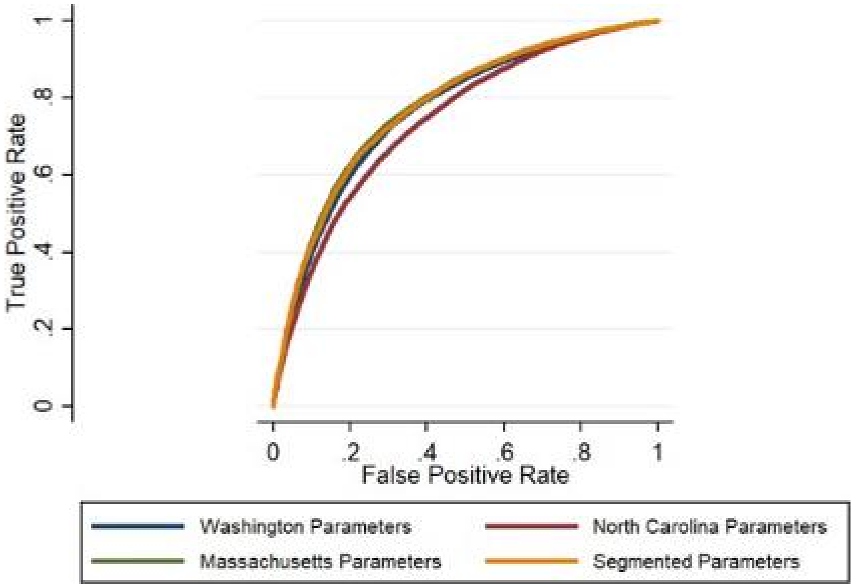


Figure A20: ROC Curve of Graduation by Model Specification in Washington

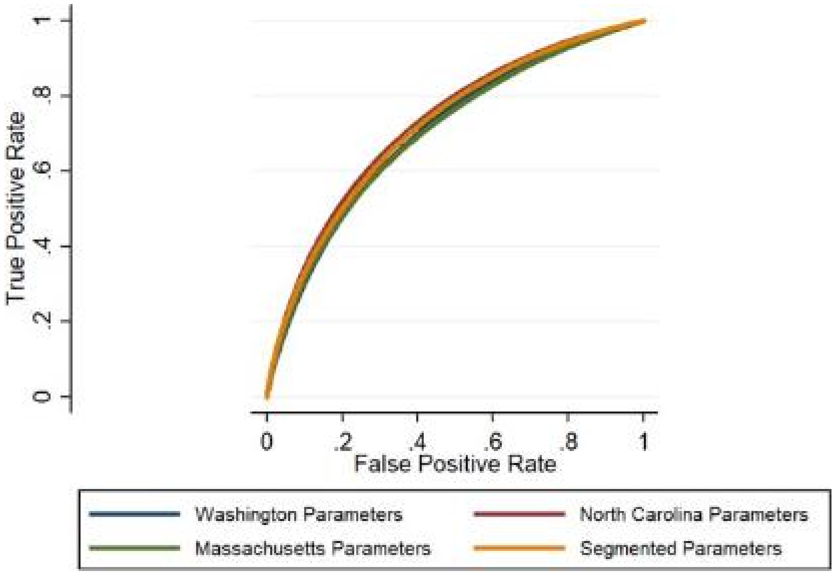


Figure A21: ROC Curve of Graduation by Model Specification in North Carolina

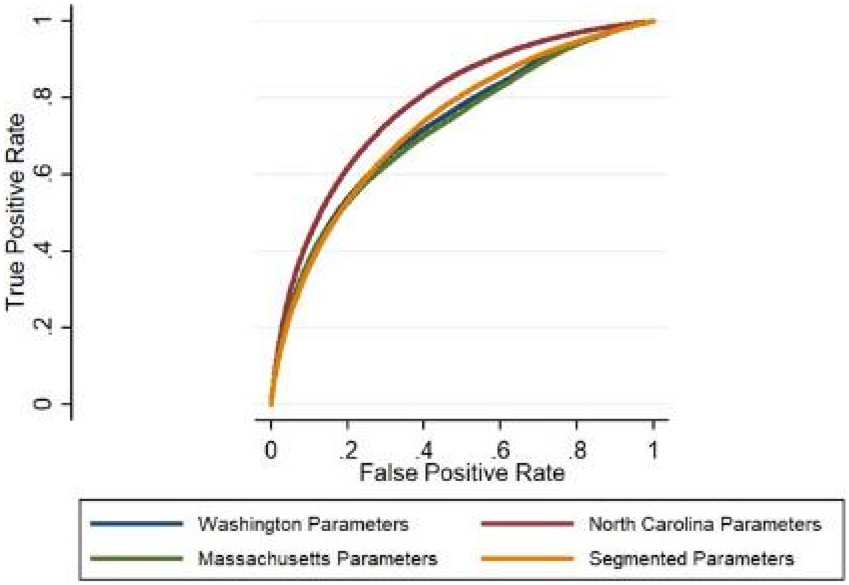


Figure A22: Probability of Graduation by 3rd Grade Test Score Decile and EDS in Washington

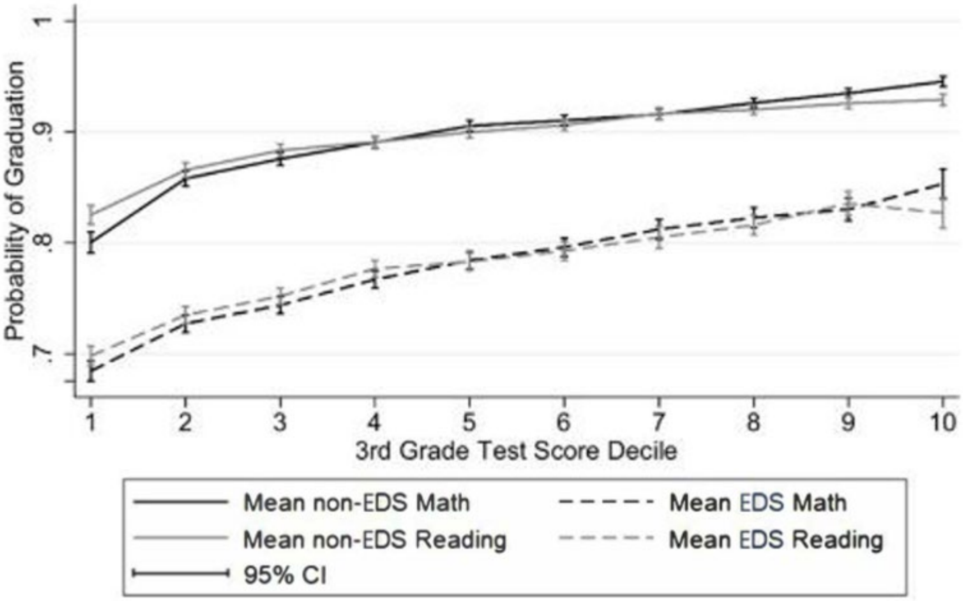


Figure A23: Probability of Graduation by 3rd Grade Test Score Decile and EDS in Massachusetts

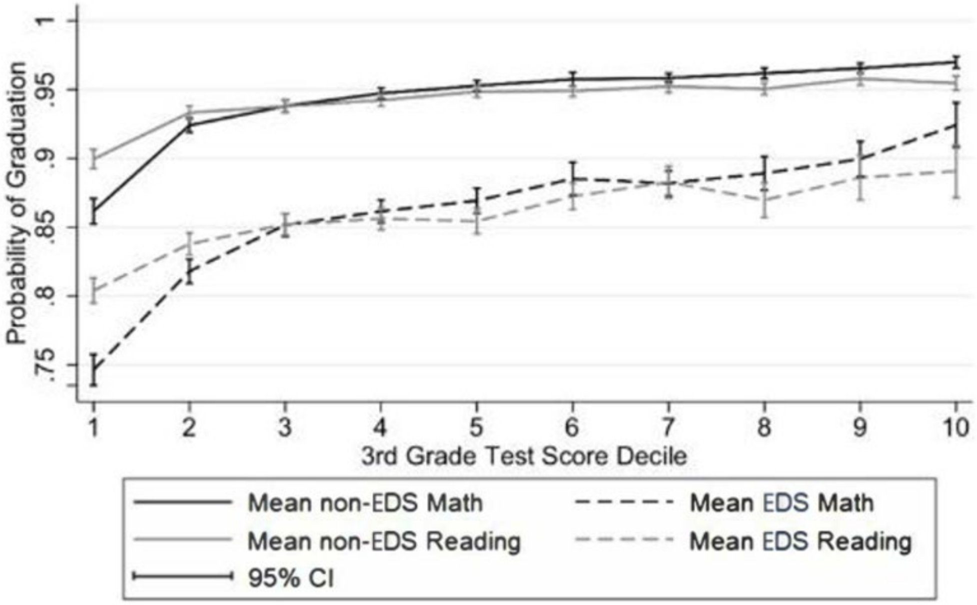


Figure A24: Probability of Graduation by 3rd Grade Test Score Decile and EDS in North Carolina

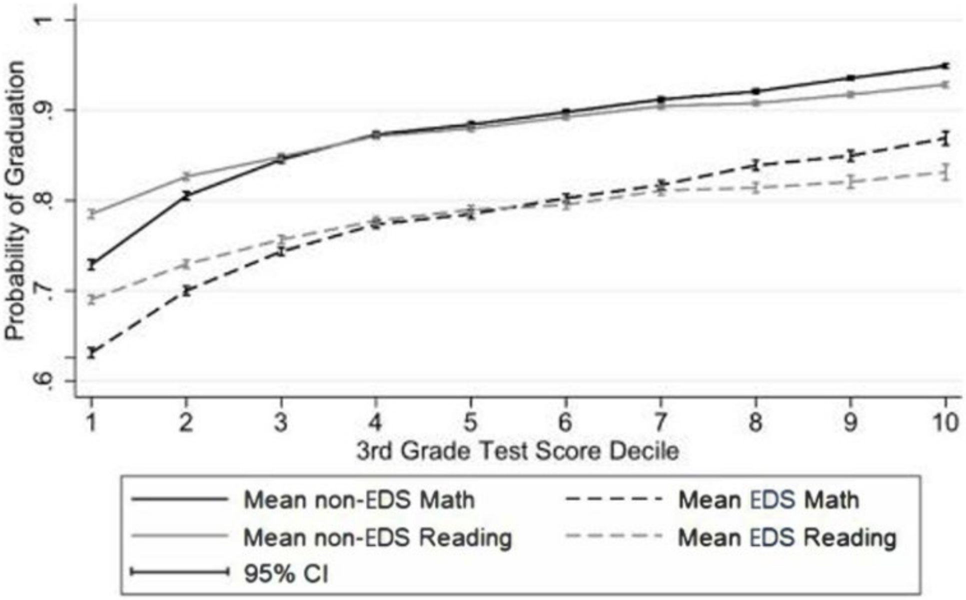


Figure A25: Probability of Advanced Course-Taking by 3rd Grade Test Score Decile and EDS in Washington

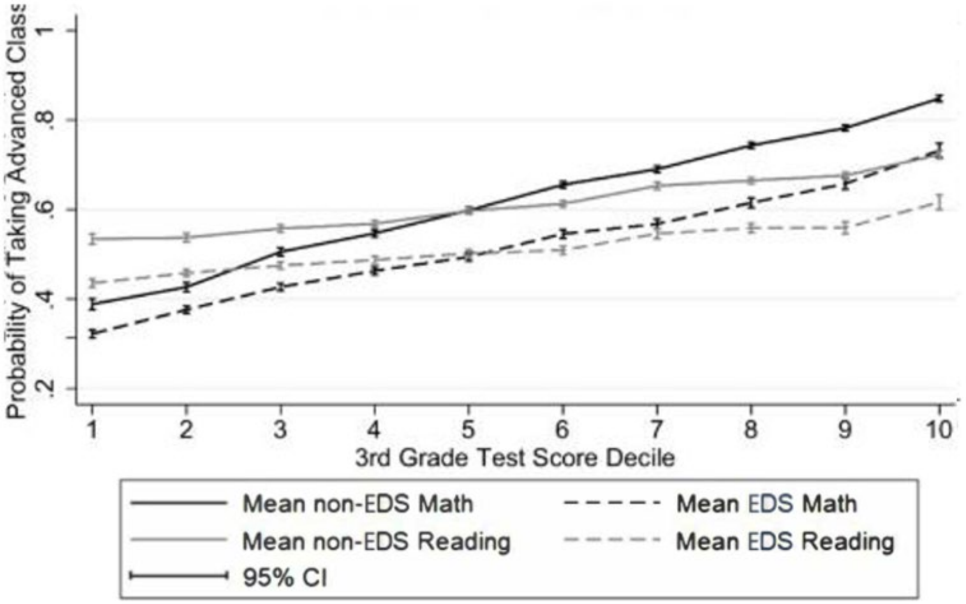


Figure A26: Probability of Advanced Course-Taking by 3rd Grade Test Score Decile and EDS in Massachusetts

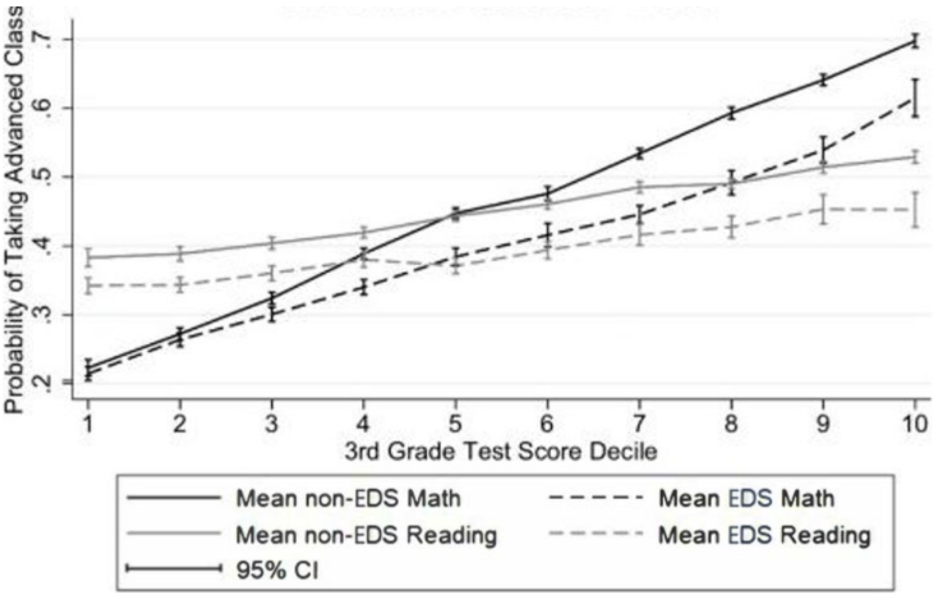


Figure A27: Probability of Advanced Course-Taking by 3rd Grade Test Score Decile and EDS in North Carolina

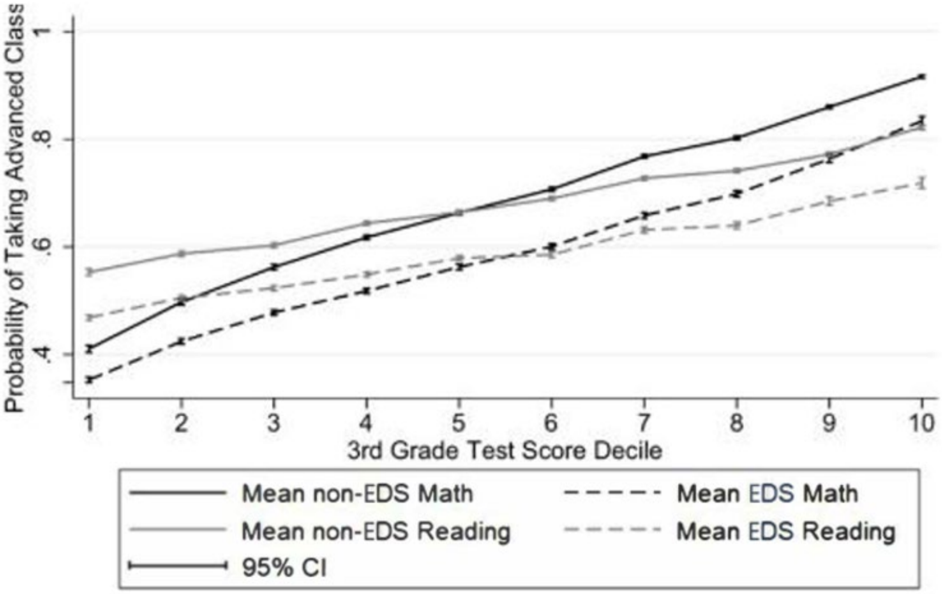


Figure A28: High School Math Percentile by 3rd Grade Test Score Decile and EDS in Washington

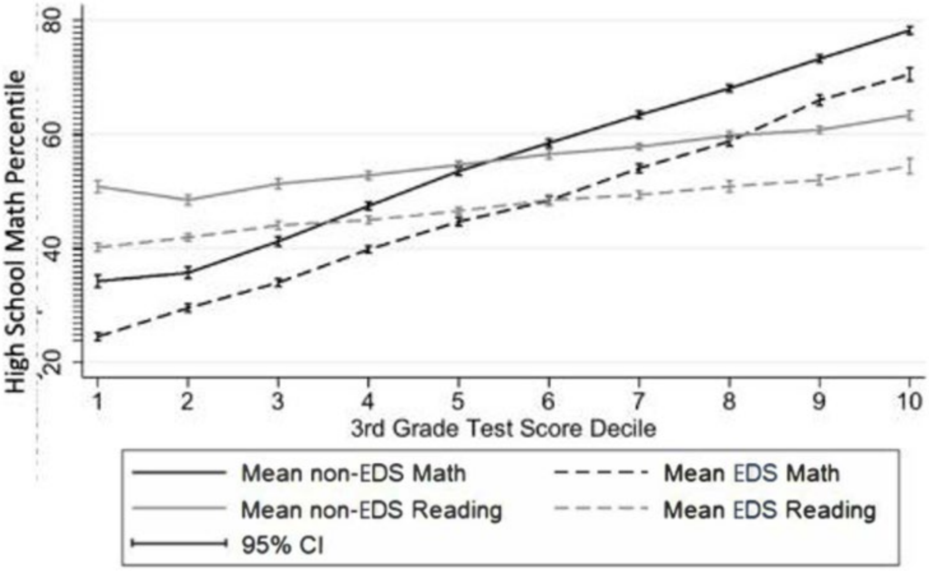


Figure A29: High School Math Percentile by 3rd Grade Test Score Decile and EDS in Massachusetts

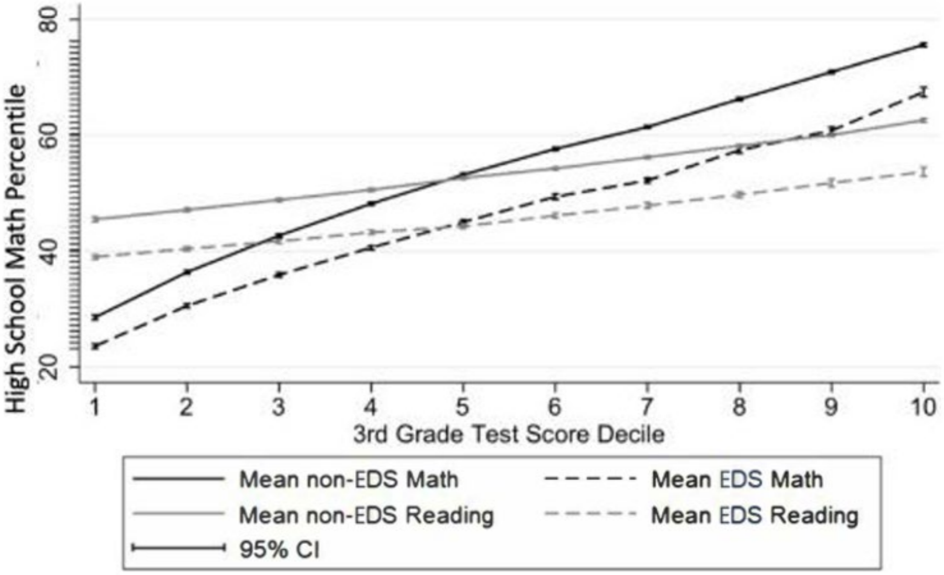


Figure A30: High School Math Percentile by 3rd Grade Test Score Decile and EDS in NorthCarolina

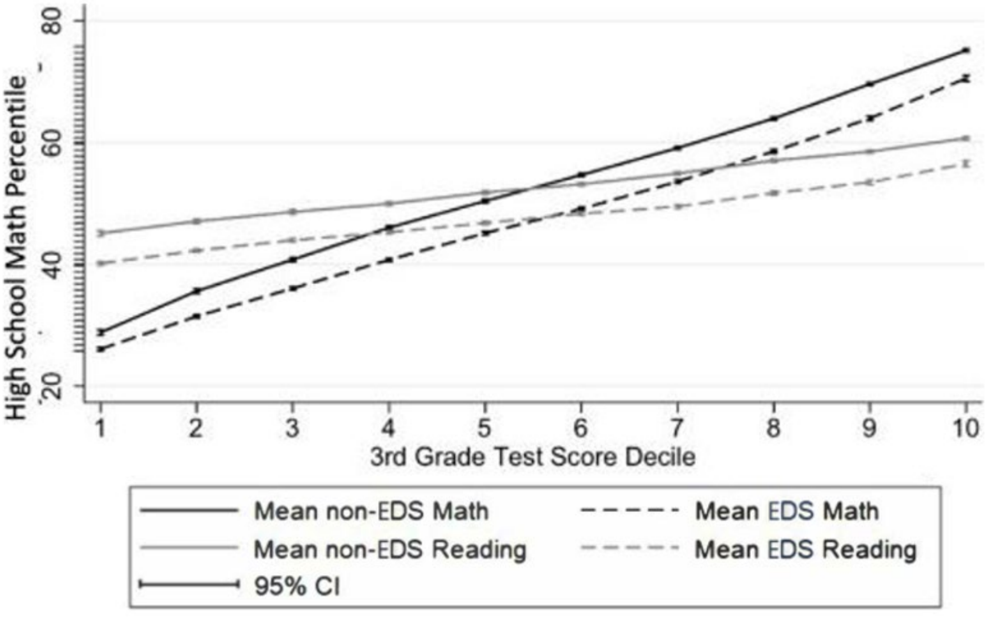


Figure A31: Probability of Top 50th Percentile in 8th Grade Math by 3rd Grade Test Scores and EDS in Washington

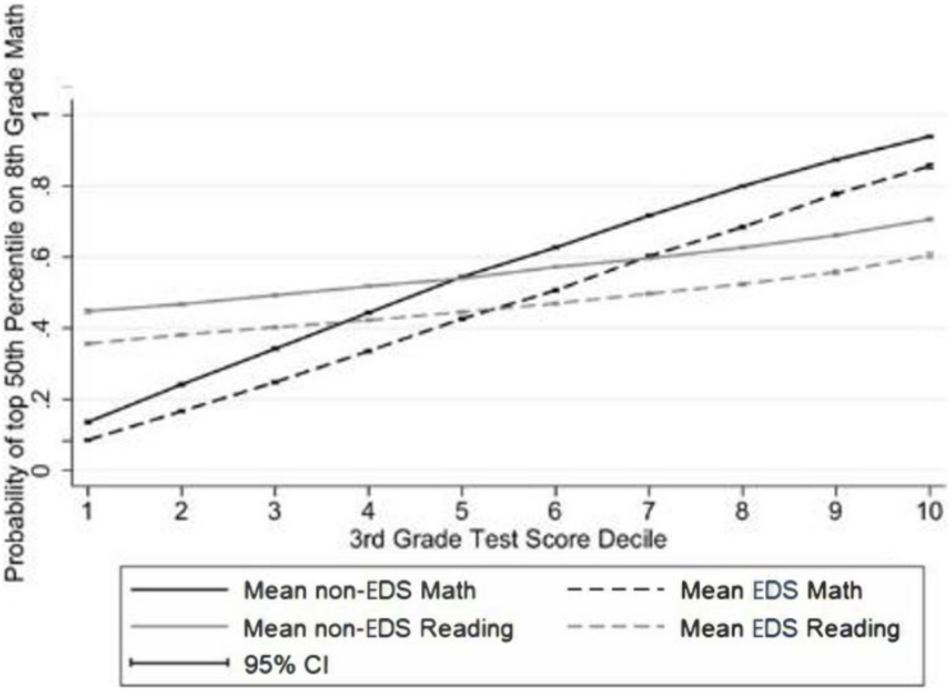
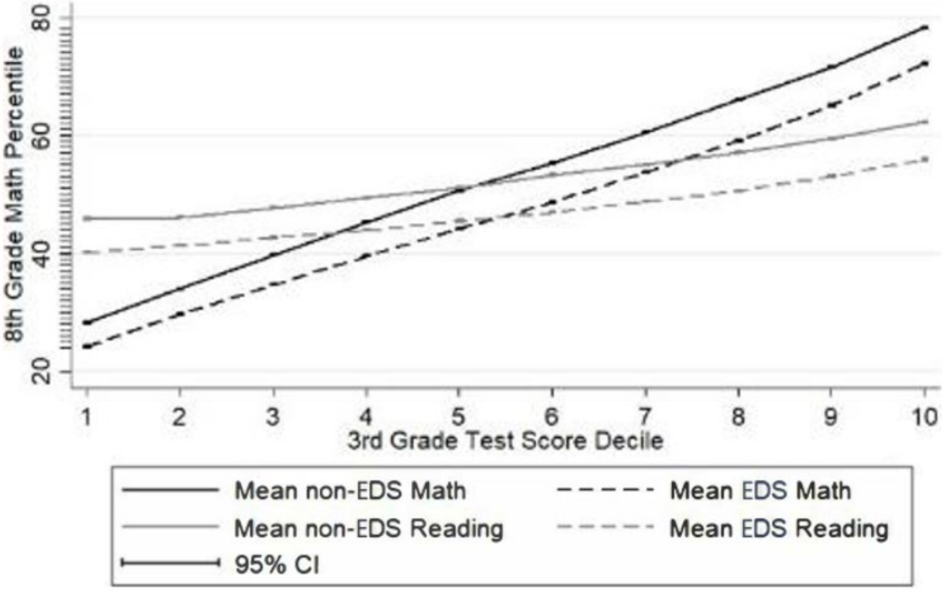


Figure A32: 8th Grade Math Percentile by 3rd Grade Test Scores and EDS in Massachusetts



Appendix B

Table B1: Model Coefficients for Additional Outcomes by State

Panel A: 8th Grade Testing Distribution

	Overall	MA	NC	WA
	(A1)	(A2)	(A3)	(A4)
3rd Grade Math Percentile	0.535***	0.494***	0.558***	0.508***
	(0.000709)	(0.00162)	(0.000921)	(0.00154)
3rd Grade Reading Percentile	0.179***	0.205***	0.169***	0.180***
	(0.000722)	(0.00166)	(0.000933)	(0.00158)
N	2,014,604	382,772	1,213,361	418,471

Panel B: Probability Top Half of the 8th Grade Testing Distribution

	Overall	MA	NC	WA
	(B1)	(B2)	(B3)	(B4)
3rd Grade Math Percentile	0.617***	0.558***	0.643***	0.603***
	(0.00120)	(0.00284)	(0.00155)	(0.00261)
3rd Grade Reading Percentile	0.185***	0.211***	0.168***	0.199***
	(0.00134)	(0.00312)	(0.00172)	(0.00294)
N	2,014,604	382,772	1,213,361	418,471

Panel C: Probability Top Half of the High School Testing Distribution

	Overall	MA	NC	WA
	(C1)	(C2)	(C3)	(C4)
3rd Grade Math Percentile	0.570***	0.577***	0.566***	0.614***
	(0.00192)	(0.00329)	(0.00253)	(0.00656)
3rd Grade Reading Percentile	0.168***	0.192***	0.163***	0.175***
	(0.00211)	(0.00363)	(0.00276)	(0.00741)
N	824,324	285,396	480,682	58,246

Notes: All models are estimated using linear regression. Columns (1) display no imputation, columns (2) display standard imputation described in Section 5.3, and columns (3)-(6) display imputation with ad hoc adjustments to test score coefficients described in section 5.3. The regression sample includes students who have 3rd grade math and reading test scores and 3rd grade student characteristics. All regressions control year, student race, gender, ethnicity,disability status, English language learner status, economically disadvantaged status, and enrollment status in specialeducation.

* *p<0.10* ** *p<0.05* *** *p<0.01.* Probability values are from a two-sided *t*-test.

Table B2: Model Coefficients of 8th Grade Math Percentile by State and Imputation ValuePanel A: Massachusetts

	Not Imputed	Imputed	+10%	+25%	-10%	-25%
	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)
3rd Grade Math Percentile	0.503***	0.513***	0.514***	0.516***	0.511***	0.509***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
3rd Grade Reading Percentile	0.210***	0.211***	0.211***	0.212***	0.211***	0.210***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
R Squared	0.550	0.584	0.585	0.587	0.582	0.578
N	382,772	482,264	482,264	482,264	482,264	482,264

Panel B: North Carolina

	Not Imputed	Imputed	+10%	+25%	-10%	-25%
	(B1)	(B2)	(B3)	(B4)	(B5)	(B6)
3rd Grade Math Percentile	0.563***	0.572***	0.574***	0.576***	0.570***	0.565***
	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
3rd Grade Reading Percentile	0.170***	0.176***	0.176***	0.177***	0.175***	0.174***
	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
R Squared	0.580	0.619	0.621	0.623	0.616	0.611
N	1,213,361	1,505,484	1,505,484	1,505,484	1,505,484	1,505,484

Panel C: Washington

	Not Imputed	Imputed	+10%	+25%	-10%	-25%
	(C1)	(C2)	(C3)	(C4)	(C5)	(C6)
3rd Grade Math Percentile	0.510***	0.518***	0.519***	0.521***	0.517***	0.514***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
3rd Grade Reading Percentile	0.180***	0.183***	0.184***	0.184***	0.183***	0.182***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
R Squared	0.543	0.561	0.562	0.564	0.559	0.555
N	418,471	493,228	493,228	493,228	493,228	493,228

Notes: All models are estimated using linear regression. Columns (1) display no imputation, columns (2) display standard imputation described in Section 5.3, and columns (3)-(6) display imputation with ad hoc adjustments to test score coefficients described in section 5.3. The regression sample includes students who have 3rd grade math and reading test scores and 3rd grade student characteristics. All regressions control year, student race, gender, ethnicity, disability status, English language learner status, economically disadvantaged status, and enrollment status in specialeducation.

* $p<0.10$ ** $p<0.05$ *** $p<0.01$. Probability values are from a two-sided *t*-test.

Table B3: Model Coefficients by State with Imputation for High School Math Percentile (3rd Grade)

Panel A: Massachusetts

	Non Imputed	Imputed	+10%	+25%	-10%	-25%
	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)
3rd Grade Math Percentile	0.500***	0.498***	0.499***	0.500***	0.497***	0.494***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
3rd Grade Reading Percentile	0.184***	0.189***	0.189***	0.189***	0.189***	0.188***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
R Squared	0.536	0.567	0.569	0.571	0.566	0.562
N	285,396	344,462	344,462	344,462	344,462	344,462

Panel B: North Carolina

	Non Imputed	Imputed	+10%	+25%	-10%	-25%
	(B1)	(B2)	(B3)	(B4)	(B5)	(B6)
3rd Grade Math Percentile	0.487***	0.467***	0.474***	0.482***	0.459***	0.442***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
3rd Grade Reading Percentile	0.165***	0.167***	0.168***	0.170***	0.165***	0.161***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
R Squared	0.450	0.441	0.452	0.466	0.427	0.398
N	480,682	637,017	637,017	637,017	637,017	637,017

Panel C: Washington

	Non Imputed	Imputed	+10%	+25%	-10%	-25%
	(C1)	(C2)	(C3)	(C4)	(C5)	(C6)
3rd Grade Math Percentile	0.539***	0.533***	0.533***	0.533***	0.533***	0.533***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
3rd Grade Reading Percentile	0.168***	0.166***	0.166***	0.166***	0.166***	0.166***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
R Squared	0.579	0.579	0.579	0.579	0.579	0.579
N	58,246	58,246	58,246	58,246	58,246	58,246

Notes: All models are estimated using linear regression. Columns (1) display no imputation, columns (2) display standard imputation described in Section 5.3, and columns (3)-(6) display imputation with ad hoc adjustments to test score coefficients described in section 5.3. The regression sample includes students who have 3rd grade math and reading test scores and 3rd grade student characteristics. All regressions control year, student race, gender, ethnicity, disability status, English language learner status, economically disadvantaged status, and enrollment status in specialeducation.

* *p<0.10* ** *p<0.05* *** *p<0.01.* Probability values are from a two-sided *t*-test.

Table B4: Model Coefficients by State with Imputation for Advanced Course-Taking (3rd Grade)

Panel A: Massachusetts

	Non Imputed	Imputed	+10%	+25%	-10%	-25%
	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)
3rd Grade Math Percentile	0.528***	0.528***	0.528***	0.528***	0.528***	0.528***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
3rd Grade Reading Percentile	0.188***	0.187***	0.187***	0.187***	0.187***	0.187***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
R Squared	0.190	0.190	0.191	0.191	0.191	0.191
N	172,243	172,651	172,651	172,651	172,651	172,651

Panel B: North Carolina

	Non Imputed	Imputed	+10%	+25%	-10%	-25%
	(B1)	(B2)	(B3)	(B4)	(B5)	(B6)
3rd Grade Math Percentile	0.478***	0.470***	0.478***	0.478***	0.479***	0.479***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
3rd Grade Reading Percentile	0.260***	0.253***	0.263***	0.262***	0.264***	0.265***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
R Squared	0.211	0.208	0.216	0.215	0.216	0.216
N	773,644	787,543	787,543	787,543	787,543	787,543

Panel C: Washington

	Non Imputed	Imputed	+10%	+25%	-10%	-25%
	(C1)	(C2)	(C3)	(C4)	(C5)	(C6)
3rd Grade Math Percentile	0.467***	0.464***	0.467***	0.467***	0.467***	0.466***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
3rd Grade Reading Percentile	0.205***	0.204***	0.206***	0.206***	0.206***	0.205***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
R Squared	0.176	0.175	0.178	0.178	0.178	0.177
N	242,333	244,964	244,964	244,964	244,964	244,964

Notes: All models are estimated using linear regression. Columns (1) display no imputation, columns (2) display standard imputation described in Section 5.3, and columns (3)-(6) display imputation with ad hoc adjustments to test score coefficients described in section 5.3. The regression sample includes students who have 3rd grade math and reading test scores and 3rd grade student characteristics. All regressions control year, student race, gender, ethnicity, disability status, English language learner status, economically disadvantaged status, and enrollment status in specialeducation.

* *p<0.10* ** *p<0.05* *** *p<0.01.* Probability values are from a two-sided *t*-test.

Table B5: Model Coefficients by State with Imputation for Graduation (3rd Grade)

Panel A: Massachusetts

	Non Imputed	Imputed	+10%	+25%	-10%	-25%
	(A1)	(A2)	(A3)	(A4)	(A5)	(A6)
3rd Grade Math Percentile	0.105***	0.0846***	0.102***	0.102***	0.101***	0.0996***
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
3rd Grade Reading Percentile	0.054***	0.046***	0.052***	0.053***	0.051***	0.050***
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
R Squared	0.093	0.071	0.108	0.107	0.108	0.107
N	172,651	207,520	207,520	207,520	207,520	207,520

Panel B: North Carolina

	Non Imputed	Imputed	+10%	+25%	-10%	-25%
	(B1)	(B2)	(B3)	(B4)	(B5)	(B6)
3rd Grade Math Percentile	0.221***	0.146***	0.215***	0.215***	0.211***	0.207***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
3rd Grade Reading Percentile	0.100***	0.106***	0.141***	0.142***	0.136***	0.132***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
R Squared	0.104	0.065	0.131	0.131	0.129	0.124
N	786,564	1,069,956	1,069,956	1,069,956	1,069,956	1,069,956

Panel C: Washington

	Not Imputed	Imputed	+10%	+25%	-10%	-25%
	(C1)	(C2)	(C3)	(C4)	(C5)	
	0.139***	0.114***	0.001***	0.001***	0.001***	0.001***
3rd Grade Math Percentile	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
	0.115***	0.095***	0.114***	0.114***	0.113***	0.111***
3rd Grade Reading Percentile	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
R Squared	0.083	0.063	0.092	0.092	0.092	0.092
N	244,964	278,690	278,690	278,690	278,690	278,690

Notes: All models are estimated using linear regression. Columns (1) display no imputation, columns (2) display standard imputation described in Section 5.3, and columns (3)-(6) display imputation with ad hoc adjustments to test score coefficients described in section 5.3. The regression sample includes students who have 3rd grade math and reading test scores and 3rd grade student characteristics. All regressions control year, student race, gender, ethnicity,disability status, English language learner status, economically disadvantaged status, and enrollment status in specialeducation.

* *p<0.10*   ** *p<0.05*   *** *p<0.01*. Probability values are from a two-sided *t*-test.