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

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 • May 2020

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 (grant number: G-19-56639). We would also like to thank Nathaniel Marcuson for his excellent research assistance and David Keeling for editorial suggestions. 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 • American Institutes for Research
1000 Thomas Jefferson Street NW, Washington, DC 20007
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
May 2020

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 intended to be used as a diagnostic tool to identify individual student needs, so that students can be set on a trajectory for long-term academic success. We use panel data from three states – North Carolina, Massachusetts and Washington State – to investigate how accurate early measures of achievement are in predicting later high school outcomes. We contribute to the literature in three distinct ways. First, the long panels we employ allow us to quantify the accuracy of models predicting how early (3rd and 4th grade) measures of student background and achievement predict several later schooling outcomes: 8th grade 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. grades 3 to 8, then 8 to 12) sacrifice forecast accuracy; this is of particular policy relevance for states or localities that do not yet have long administrative data panels. Finally, 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 other states.

1. Introduction

Testing students annually and using the results to inform policy decisions, including school accountability, has been one of the primary federal strategies for identifying and addressing educational inequity over the past decade and a half. But tests are also intended to be useful as informative and diagnostic tools for educators and parents. Test results hold the potential to identify individual student needs, so that students can receive needed help on a timely basis. Both Chicago (Allensworth, 2013) and Massachusetts (Jung et al., 2012), for example, have developed early warning systems that use administrative data to predict whether students are at-risk of not meeting specified academic outcomes. However, existing early warning systems are somewhat limited, relying primarily on data from middle school grades.[1] Even when students, parents, and educators receive an "early" warning, there is very little time to intervene before students enter high school. We investigate the accuracy of warnings that could be generated much earlier in a student's career – beginning in third grade.

Past 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 now changing. All states since 2005-06 have been testing students annually in math and reading in grades three through eight as a result of the No Child Left Behind Act passed in 2002 (Le Floch et al., 2007). Thus, a number of state administrative datasets now have the capacity to follow students over their entire K-12 schooling careers (and in some cases beyond), permitting predictions of long-term student achievement, such as the probability of high school graduation, based on information about students from their elementary years.

While there is disagreement about the use of tests for accountability and/or the amount of testing (e.g., Goldhaber and Ozek, 2019; Koretz, 2017), there remains broad support for the general idea of using data to help understand and inform stakeholders on student progress and educational outcomes. For example, a recent national poll of parents and teachers released by the Data Quality Campaign (September, 2019) revealed that 90% of parents say they need data to understand their child's progress and help them do their best, and 86% of teachers believed using data (e.g. for planning instruction, identifying learning goals of students, or knowing what concepts students are learning) is an important part of being an effective teacher.

Test results can provide objective information to clarify perceptions about student progress but there are often striking differences between teacher and parental perceptions of test performance and the implications for this. For instance, according to research by Learning Heroes (October 2019), 90% of parents believe their child is performing at or above grade level in math and reading. By contrast, surveys of teachers tend to find far fewer believe that students

[1] In Chicago, for instance, researchers use 8th grade test scores, grades, and attendance to predict the likelihood that students will graduate and provide high schools with a list of their incoming students at risk of failure. The system in Massachusetts is similar, though their early warning system covers different grade spans and outcomes. Thus, neither of these systems fully exploits the long panels of data that now exist in many states, e.g. using early elementary indicators to assess risk of dropping out of high school.

are academically on track; a survey by Scholastic (2016), for instance, finds just 39% of teachers reporting their students come prepared for grade level work at the start of the school year.

Predictions of long-term student outcomes are useful only if they are reasonably accurate. While the potential for data to inform parents about student progress is clear, there is also legitimate concern about the harm that might occur if the information is inaccurate or misleading.[2] 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 three 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., grades 3 to 8, then 8 to 12) sacrifice forecast accuracy (which is of particular policy relevance for states or localities that do not yet have long administrative data panels). Finally, 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 other states.

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 68% 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.

We also find that predictive models "travel" across state lines effectively. 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 72% 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-

[2] In the case of school decisions, this could include poor allocation of school resource, 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. 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 (2013), 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.

graduates are misclassified.[3] 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, being a free or reduced-price lunch student in 3rd grade lowers the student's predicted position in the high school math distribution by 6 percentile points, the predicted probability of taking an advanced course in high school by 8.4 percentile points, and the predicted probability of graduation by 10 percentage points. Put another way, a student receiving free or reduced-price lunch 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-free or reduced-price lunch student 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.[4] 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.[5] 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, and the risk is highest for

[3] In other words, with a misclassification rate of 28%, cross-state estimates correctly classify as little as 2% fewer graduates compared to in-state estimates.

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

[5] 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.

the lowest performers, and this effect is even more pronounced for free and reduced price lunch students.[6] This widely-cited report was influential in shaping federal and state early reading intervention strategies; at least four states have passed third grade reading laws since the report's release,[7] while other states have amended their third grade reading laws multiple times or phased in various requirements (CCSSO, 2019).

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 course-taking patterns, in math and science. And two other recent studies show that students' high school GPA is a strong predictor of high school graduation (Easton, Johnson, and Sartian, 2017), and college-going, retention, and graduation (Geiser and Stanelices, 2007; Easton, Johnson, and Sartian, 2017).[8] Similarly, Silver et al. (2008) follow a cohort of Los Angeles Unified School District students over a seven year period, find that test scores as early as grade 6 are predictive of on-time graduation.

Zau and Betts (2008) addresses 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,[9] the authors find evidence suggesting predictions using 4th grade test scores and student characteristics have nearly the same accuracy as predictions when 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",[10] 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, Balfanz, and Herzog 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, and Allensworth and Easton (2005) find that students having accumulated 5 full course credits with no more than one

[6] 23% of the lowest performing readers do not graduate high school on time, relative to 9% for "basic" readers and 4% 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.

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

[8] 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.

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

[10] 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% passed in the following year.

semester F in a core subject is more indicative of high school graduation than standardized test scores for Chicago Public Schools (CPS) students.[11] While these results do not control for other factors and are ultimately based on summary statistics of relatively small cohorts, they indicate a complex dependence structure between early student characteristics and long-term outcomes that can't be solely captured by student testing.

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. With over 6 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 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 8% of non-graduates, boosted decision trees increases the correct classification of graduates by approximately 20% at the same misclassification rate.

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 personnel 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 (EWIS) 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 in 2011 released a state-wide Early Warning Indicator Index (EWII) using academic and behavioral student characteristics[12] 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 extent to which a student's rank in the distribution of academic performance changes during their schooling career. Using test score

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

[12] Student characteristics include spring 2011 grade 8 Massachusetts Comprehensive Assessment Exam (MCAS) results, spring 2011 grade 8 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 1, 2011.

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.[13]

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, examines the degree to which early test scores predict advanced course-taking, the degree to 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 2013, where depending on the outcome and state contain as many as 16 cohorts of students.[14] 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 (SCS), and demographic information and high school exit codes from the Student Information and Management System (SIMS), all of which are provided by the Massachusetts Department of Education (DOE). The North Carolina longitudinal student data combines annual North Carolina Education Research Data Center (NCERDC) End-of-Grade files, Masterbuild files, and AP course membership files, which include student-level characteristics such as URM status, FRL status, AP course taking behavior and high school exit codes. The longitudinal student data in Washington combines the state's Core Student Records System (CSRS) and Comprehensive Education Data and Research System (CEDARS), both

[13] 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.

[14] 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.

maintained by OSPI, which detail student-level characteristics such as URM and FRL status and high school exit codes as well as AP course membership.

This panel allows us to leverage students' standardized tests throughout their academic career, and link these to their course-taking patterns in high school, and identify time of graduation. However, there are two particular caveats precluding generalization to the entire student population. First, the high school testing outcomes of interest are inconsistent across states and time (we describe these in greater detail in **Appendix B**). While there is some evidence that test regime changes generally have small impacts on placement in the test distribution (Backes et al., 2018) for end-of-year tests, there is little evidence about whether regime changes that lead to tests taken in different grades or for more focused subjects may affect the degree to which earlier tests are predictive. This is a concern given that both types of changes occur in our sample. For instance, the majority of students in Massachusetts take a high school math test in 10th grade, whereas the grade for the majority of students taking a math test in North Carolina and Washington varies 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 type.[15]

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. 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.[16]

To provide a sense of the nature of attrition we show, in **Figure 1**, the average percentage of 3rd grade cohorts who are observed in the subsequent grades by grade and by state. 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 to 30 percent 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,[17] but detailed exploration of differences in sample attrition across states is outside the scope of this study.

---

[15] 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.

[16] 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.

[17] 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. It's outside the scope of our analysis, it may be that there are state-to-state differences in private school enrollment or homeschooling.

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. As Austin et al. (2020) describe, there are two distinct issues. The first is "reference bias," that is the degree to which the attrition from the sample changes the distributions we are using for comparison. This is only an issue for the high school test outcome as advanced course-taking and high school graduation are independent of the achievement of those who leave the state samples.[18] If attrition is solely related to observable attributes then reference bias is limited to an issue of generalizability, i.e. we could not generalize the findings outside of the students that remain in the sample. But this is problematic for the potential uses of making long-term projects of student achievement as we would like to be able to draw inferences about the likely achievement trajectories of a random 3rd grader, not one who stays in a state data system. Thus, to account for the reference bias issue, consistent with Austin et al. (2020), we "backfill" the achievement distribution by normalizing later test outcomes using *all* students in a state, not just those who were in the 3rd grade sample and who had a later high school outcome.[19]

A second issue of bias arises 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.[20] Such a relationship would lead to bias in the estimated parameters of the model (we describe how we address the potential issues associated with sample selection in Section 4.1).

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.

We present cohort sizes of the analytic sample by outcome, state, and school year in Appendix A **Table A1)**.[21] North Carolina has up to 14 cohorts and has the largest cohorts, with an average cohort size of 72,000 for graduation outcomes, 70,000 using advanced course-taking

[18] Assuming that who ends up in advanced courses and graduates from high school are not affected by high school exiters.

[19] Austin et al. (2020) report that while there is some heterogeneity of in- and out-of-state-sample migrators, the achievement of students coming into state samples after 3rd grade is similar to the achievement of students who leave the samples.

[20] 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.

[21] We represent school years by their right endpoint, e.g. the schoolyear 1997-1998 is represented as 1998.

outcomes, and 80,000 using high school testing outcomes. The panels in Massachusetts (5 cohorts) and Washington[22] (5 cohorts) are similar with about 57,000 to 61,000 for the various outcomes. The specific sample sizes for each of the later high school outcomes we observe, for 3rd graders, are provided in **Table A1**. 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 3rd graders in 2010-2011, the last observable year in our 3rd grade panel.

Table 1 provides 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 (28%) than other two states (5%-8%), and a notably smaller population of Asian and Pacific Islander students in North Carolina (2%) than other states (6-9%).

Graduation rates for 3rd graders that we track to high school are relatively similar across the states, in the range of 80-90 percent across individual cohorts.[23] Advanced course-taking varies significantly more: while 55% of students in the sample take at least one advanced math or science course in Washington, and 62% in North Carolina, only 37% of students in Massachusetts are identified as taking an advanced math or science course. These differences are likely based on the definitions of the advanced course-taking measures. 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, across all states we see similar patterns based on the quartile of 3rd grade math test achievement: lower test scoring students tend to be disproportionately represented by various measures of student disadvantage (being eligible for free or reduced price lunch, being identified as limited English proficient or with a learning disability or special needs, being in an

[22] 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.

[23] 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/), 87% for North Carolina in 2017 (Newsroom, 2017), and 79.3% for Washington in 2016 (Weaver-Rendall & Ireland, 2018).

underrepresented minority category). And, also 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 follow the approach in Austin et al. (2020), estimating linear models where the ranking of students in the high school math test distribution is a function of 3rd grade ranking and observable student characteristics:

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

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, free or reduced-price lunch (FRL) 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. Our primary focus is on β_1, which indicates the relationship between 3rd grade tests and high school outcomes (high school math tests in (1)).[24] We also estimate several extensions of the above model, including replacing or supplementing 3rd grade test rankings with 8th grade rankings on math and ELA tests, estimating the various specifications separately by state, and estimating models with 8th grade math score percentile as an outcome.[25]

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

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

$$Z_i = \beta_0 + T_i'\beta_1 + X_i'\beta_2 + \varsigma_i T_i'\beta_3 + \varsigma_i + \delta_i + \varepsilon_i, \qquad (2b)$$

[24] 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.

[25] The estimation of separate state models is important for testing the degree to which estimates from one state can reliably be used for generating predictions of student achievement in a different state.

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 an 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.

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. Since, in general, out-of-sample contexts will not have access to state or cohort fixed effects, we estimate the above models without state and 3rd grade cohort fixed effects to assess the predictive accuracy of equations (1) and (2). Secondly, we define a metric below more purposefully related to the classifications of students into categories. This 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. It is also important for the analysis (described in more detail in 4.2) where we are analyzing the efficacy of using cross-state models and segments of achievement data (3rd to 8th then 8th to 10th) to make long-term outcome predictions.

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. There is, however, significant prior evidence that student background, such as low-income status, is correlated with long-term academic performance (e.g., Goldhaber et al., 2018; Reardon, 2011). 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 more flexible specifications that models in which we include decile of test score achievement in 3rd grade and interactions between these deciles and student characteristics.

Finally, as noted above, there is concern that sample attrition could lead to biased estimates of the relationship between 3rd grade test achievement and high school outcomes through two primary mechanisms. First, students may be differentially likely to leave the sample based on their placement on the testing distribution. Second, students who leave the sample may have a different relationship between their early test scores and long-term outcomes than those who stay. We address the issue of these potential sources of sample selection bias in two ways: assessing model coefficients and predicted outcomes with subsets of the data that are defined based on sample attrition at different points in students' academic careers and imputing outcomes using ad hoc adjustments to test score effects for leavers to bound the potential bias arising from a differential relationship between early test scores and high school outcomes. These approaches are described in greater detail in Section 5.3, but, as it turns out, they have end up resulting in little change in the estimates.

4.2 Evaluation of Predictive Power

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 predicted probabilities $\hat{p}_i$ relative to dichotomous outcome variables is less straightforward. We focus on 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),[26] we alleviate this issue by 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.[27]

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 you pick, 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.

[26] 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.

[27] 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.

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

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

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.

4.3 Measuring the Accuracy of Out-Of-Sample Predictions

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.[28] 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, 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, Tibshirani, & Friedman, 2005). Specifically, we are interested in estimating the expected "test error" TE:

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

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 $K_1, \ldots, K_{10}$ (or, more generally, κ partitions) and calculate the expected test error

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

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

where $(\tilde{X}_l, \tilde{Y}_l)$ is the subset of $(\tilde{X}, \tilde{Y})$ in partition K_l and $\hat{p}_{K_l}$ is estimated on the set $\{K_i \,|i \neq l\}$. 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,

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

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:

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

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:

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

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.[29] Let the superscripts ρ_i denote distinct panels of students. We first estimate the relationship between 3rd 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):

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

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

$$\widehat{Y_\iota} = g\left(\widehat{\beta_0} + \widehat{Z_\iota}\,\widehat{\beta_1} + \varepsilon_i\right), \tag{11}$$

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

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 2007-2009 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.[30]

[29] For instance, the length of administrative panels in some 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 3rd 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.

[30] 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.

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 in high school, and the probability of graduating high school.[31] 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 (this is discussed in greater detail in Section 5.2).[32]

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 quite strong correlations between a student's place in the distributions of 3rd grade and high school math tests (column 1), 8th grade and high school tests (column 2), and we also find that both 3rd and 8th grade tests (column 3) are independently predictive of a student's place in the high school test distribution.[33] This last finding suggests that the trajectory of achievement from 3rd grade to 8th grade matters in the case of predicting high school math achievement, and as shown in **Table 4**,[34] correlations of predicted high school math percentiles across states and grade are above .79.

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 FRL, etc.) predicted to be significantly lower in the high school test distribution. Relative to non-FRL students, for instance, FRL students are predicted to 2 to 6 percentile points lower in the high school test distribution.

Columns 4-6 and 7-9 present analogous specifications for the probability of advanced course-taking and high school graduation, respectively. The trends in the coefficients on base

[31] 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.

[32] The likelihood ratio test for each of the dichotomous high school outcomes is over 87.

[33] Not surprisingly, while both math and reading tests are statistically significant in these models, the estimated relationship between base year math tests and high school math outcomes is much stronger than the relationship between base year reading tests and high school tests.

[34] This is true more generally-we find that student test scores in various combinations of grades from 3rd through 8th are independently predictive of high school outcomes, and moreover, in what order students achieve certain test scores matters. For example, 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 79%. Put another way, the risk of not graduating is 40% larger for the student with declining percentile ranking. Trajectory similarly influences advanced 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. These results are available from authors upon request.

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, as was the case with high school test percentile as the outcome, both 3rd and 8th 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 effected by whether they are generated using 3rd or 8th grade tests.

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 FRL classification to assess whether the relationships between test scores and high school outcomes differ based on their income status. We report coefficients by state in **Table 4**.

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 is 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.[35] All else equal, a student at the lowest percentile of the 3rd grade math test distribution rather than the highest percentile is expected to be 48-54 (depending on state) percentile points lower in the high school math test distribution, is expected to be 45 to 50 percent less likely to take an advanced course in high school, and 11 to 21 percent less likely to graduate.

Figures 2 to **4** show marginal effects of high school outcome by FRL status and test score decile. [36] Specifically, by each decile of math or reading and by FRL 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 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.[37,38]

These figures also highlight consistent, large effects of FRL status on predictions. Specifically, within test score decile FRL students are 8%-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, FRL recipients are 7% to 12% less likely to take an advanced course depending on test score decile and positively correlated with test score, lowering

[35] These results hold when predicting 8th grade math percentiles, see Appendix A **Table A4**.
[36] As noted above, there are also differences by other characteristics, such as student race/ethnicity, but we highlight the differences by FRL status because they tend to be much larger than those for the other student sub-groups.
[37] 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 **Figures A12**-**A20**.
[38] Large effects of FRL 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 **Figures A21**-**A23**.

probability of taking an advanced course in high school as little as a few percent at the bottom of the testing distribution, up to 8% at the top of the distribution. Finally, FRL students consistently score 3 to 5 percentile points lower on the high school math distribution.[39,40]

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. The similarity of AUC—0.749 using 3rd grade student measures versus 0.764 using 8th grade student measures shows a high predictive power and hence strong potential for effective intervention early in students' academic careers. Moreover, the findings might be interpreted as suggesting the need to intervene early as they 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 effected during schooling (Austin et al., 2020).

The difference is twice as large for advanced course-taking, with an AUC of .772 for 3rd and .802 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 9th 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 host of 9th grade characteristics, such as GPA, credit completion and number of course failures, Allensworth and Easton (2007) correctly classify 85% of graduates while misclassifying 28% of non-graduates. In comparison, our models correctly classify 68% of graduates while misclassifying 28% of non-graduates. 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, the author predicts 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 8% of non-graduates.

Figures 7 to **9** show 10-fold cross validated AUC predicting graduation and advanced course-taking, and RMSE predicting 8th grade and high school math tests using within-state

[39] We test the statistical significance of the differential nonlinear relationship by FRL status by interacting FRL status with a quadratic polynomial of test score and find that the coefficients are highly significant ($p < 0.001$).
[40] Differential effects based on non-academic characteristics are not limited to FRL 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, while American Indians are 5% less likely than white students to graduate, all else equal, and 3% less likely to take advanced courses in high school.

estimates, cross-state estimates and two-stage estimates.[41] 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 than cross-state and segmented estimates. However, the information benefit of using within state parameters is limited and statistically insignificant. For instance, in both Massachusetts and Washington the within-state estimates for high school graduation are statistically indistinguishable from both the cross-state and segmented estimates.[42,43]

Whether the cross-state or segmented estimates are more accurate varies somewhat by state and outcome. In the case of both high school graduation and advanced course-taking, differences between the cross-state and segmented estimates are statistically insignificant at the 95% confidence level for all states. The prediction approach seems to matter more for high school math test outcomes. In Massachusetts and North Carolina when predicting high school math tests the prediction accuracy of the segmented model show statistically significantly larger root mean squared error than the in-state model, but in Massachusetts and Washington, some cross-state estimates also have statistically significantly larger error than the in-state model. 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.[44]

The high correlation between model predictions across states, ranging from .835-.997 and shown in **Table 5** suggests a large amount of agreement in the relationship between 3rd grade student characteristics and high school outcomes.[45] 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

[41] 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).

[42] 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 A9-A11**. These curves illustrate the striking similarity between model specifications across the entire threshold distribution.

[43] 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 A24-A26**.

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

[45] 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.

consistently estimated relationship between student test scores across the high school math testing distribution.[46]

5.3 Assessing the Potential Implications of Sample Selection

As discussed above, we are concerned that the mobility out of state samples could lead to biased estimates of the relationship between 3rd grade test scores and high school outcomes. To understand the potential influence of sample selection bias, we regress sample attrition on 3rd grade characteristics and decile of test score. Appendix **Figures A1** to **A8** display the marginal effects of test score decile on attrition by 8th grade and by 12th grade, respectively.[47] There is significant negative selection for students moving out-of-state; the likelihood of persistence through 12th grade for students in the lowest decile of both math and reading achievement, for instance, is about 21 percentage points lower than those scoring in the middle of the testing distribution,[48] and FRL students are about 12 percentage points less likely to persist into 12th grade than their non-FRL peers.

We assess the degree of the bias in two ways. First, while we do not have true outcomes for students exiting the sample in high school, we are able to see many of their 8th grade test scores. Hence, we can get a sense of potential for sample selection bias by comparing the relationship between early academics for students who remained in the sample throughout high school and those that remained up through 8th grade.[49] We conduct regression analysis of 8th grade test scores on 3rd grade test scores with and without students who exit the sample in high school. The magnitude and direction of the difference in the resulting coefficients provide intuition on the severity and direction of potential bias. [50] As shown in Appendix **Table A1**, 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 sample, suggesting relatively little influence due to sample composition. Similarly, the difference between 3rd grade reading test percentile and 8th grade reading test percentile is .03, 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 0.98.

---

[46] 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 A27-A35**.

[47] We show marginal effects of attrition by state in Appendix A.

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

[49] This is contingent on the underlying attrition process being similar in grades 3-8 as for grades 9-12, 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.

[50] It does not, however, provide information about the effect of attrition between 3rd 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.

Second, students who exit the sample may have different relationships between early characteristics and long-term outcomes, which is of particular concern when judging the potential effects of actionable changes for students such as improving test score outcomes. We attempt to bound this potential for sample selection bias using an ad hoc imputation procedure similar to Austin et al. (2020). 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.[51] We then generate imputed values for each outcome using the most informative model available for exiters.[52] We generate baseline imputations as well as ad hoc imputations designed to bound the potential bias associated with sample attrition. 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 **Tables A5-A7**. Differences between columns (1) and (2) in these tables point to a difference in the distribution of characteristics between stayers and exiters, 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 sample attrition is not likely to be a significant source of bias in the estimated relationships.

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 lowest percentile of the 3rd grade math test distribution rather than the highest percentile is expected to be 48-54 (depending on state) percentile points lower in the high school math test distribution, is expected to be 45 to 50 percent less likely to take an advance course in high school, and 11 to 21 percent 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 changes to the predictive power of 8th grade over 3rd grade tests, suggesting that there is little

[51] 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.

[52] 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.

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 only 3rd grade student characteristics and models using 8th grade student characteristics are .89, .87, and .76 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 receiving free and reduced-price lunch in the 3rd grade in the 10th decile of the 3rd grade achievement distribution score at the level of non-FRL students in the 8th decile, are only about as likely to take an advanced math or science course in high school as non-FRL students in the 8th decile, and only about as likely to graduate as non-FRL 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 these 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.[53] However, the combination of large achievement gaps in 3rd grade and the relationship between third 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 third 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 Reardon and Jang and Reardon (2019) that explore how students' educational trajectories may vary across different contexts, and, more importantly, why it 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

[53] 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).

more students, even in settings that lack the long panels of longitudinal data included in our analysis.

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

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.

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(5)*, 1635-1664.

Burkam, D. T., & Lee, V. E. (2003). Mathematics, Foreign Language, and Science Coursetaking 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, 2009.*

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(4), 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(2), 31-47.

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(1), 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(1), 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.*

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(2), 2332858419857703.

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(2), 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(3), 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.*

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.*

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*.

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.

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

Neild, R. C., Balfanz, R., & Herzog, L. (2007). An early warning system. *Educational leadership*, 65(2), 28-33.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387), 575-583.

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

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(2), 55-64.

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

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.*

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(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(3), 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* (pp. 39-44).

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1), 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(4)*, 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

		State			Quartile of 3rd Grade Math Achievement			
	Overall	1	2	3	1 (Lowest)	2	3	4 (Highest)
Female	0.494	0.492	0.497	0.488	0.495	0.503	0.499	0.481
	(0.500)	(0.500)	(0.500)	(0.500)	(0.500)	(0.500)	(0.500)	(0.500)
Amer. Ind.	0.015	0.002	0.015	0.022	0.021	0.016	0.012	0.007
	(0.12)	(0.049)	(0.123)	(0.148)	(0.144)	(0.126)	(0.109)	(0.085)
Asian/PI	0.043	0.057	0.020	0.086	0.028	0.035	0.044	0.068
	(0.203)	(0.231)	(0.141)	(0.28)	(0.164)	(0.184)	(0.205)	(0.252)
Black	0.185	0.075	0.278	0.049	0.317	0.204	0.130	0.063
	(0.388)	(0.264)	(0.448)	(0.216)	(0.465)	(0.403)	(0.336)	(0.243)
Hispanic	0.123	0.151	0.088	0.186	0.190	0.135	0.095	0.056
	(0.329)	(0.358)	(0.283)	(0.389)	(0.392)	(0.341)	(0.294)	(0.230)
White	0.602	0.689	0.568	0.615	0.412	0.577	0.687	0.774
	(0.489)	(0.463)	(0.495)	(0.487)	(0.492)	(0.494)	(0.464)	(0.418)
Multiracial	0.04	0.026	0.045	0.040	0.043	0.042	0.040	0.035
	(0.196)	(0.160)	(0.208)	(0.195)	(0.203)	(0.201)	(0.195)	(0.184)
Lrn. Disbl.	0.051	0.057	0.047	0.052	0.106	0.047	0.025	0.010
	(0.219)	(0.233)	(0.213)	(0.223)	(0.308)	(0.211)	(0.155)	(0.100)
LEP	0.076	0.085	0.060	0.105	0.139	0.079	0.048	0.025
	(0.265)	(0.279)	(0.238)	(0.307)	(0.346)	(0.270)	(0.214)	(0.158)
FRL	0.450	0.346	0.481	0.465	0.674	0.500	0.362	0.213
	(0.497)	(0.476)	(0.500)	(0.499)	(0.469)	(0.500)	(0.481)	(0.410)
SPED	0.110	0.172	0.079	0.135	0.219	0.096	0.056	0.033
	(0.313)	(0.378)	(0.269)	(0.341)	(0.413)	(0.294)	(0.230)	(0.177)
Graduated	0.866	0.907	0.865	0.839	0.736	0.873	0.921	0.958
	(0.341)	(0.291)	(0.342)	(0.367)	(0.441)	(0.333)	(0.27)	(0.201)
Advanced	0.543	0.373	0.618	0.547	0.297	0.474	0.642	0.812
	(0.498)	(0.484)	(0.486)	(0.498)	(0.457)	(0.499)	(0.479)	(0.391)

Notes: The "Overall" column includes 3rd graders if they either had a math score or reading score, to reflect the sample for some specifications. Approximately 2% of students in the sample are missing a math score in 3rd grade

Table 2: Model Coefficients (3rd Grade)

		Highschool Math Tests			Advanced Course-Taking			Graduation		
Test Scores	3rd Grade Math Percentile	0.556***		0.155***	0.454***		0.1299***	0.166***		0.026***
		(0.004)		(0.004)	(0.002)		(0.002)	(0.002)		(0.002)
	3rd Grade Reading Percentile	0.180***		0.008**	0.212***		0.038***	0.105***		0.044***
		(0.004)		(0.004)	(0.002)		(0.002)	(0.002)		(0.002)
	8th Grade Math Percentile		0.715***	0.635***		0.593***	0.509***		0.259***	0.247***
			(0.004)	(0.004)		(0.002)	(0.002)		(0.002)	(0.002)
	8th Grade Reading Percentile		0.179***	0.143***		0.213***	0.162***		0.089***	0.07***
			(0.004)	(0.004)		(0002)	(0.002)		(0.002)	(0.002)
	R^2 or Pseudo-R^2	0.478	0.665	0.674	0.210	0.273	0.280	0.126	0.158	0.161
	N	755,751	755,751	755,751	1,110,873	1,110,873	1,110,873	1,120,023	1,120,023	1,120,023

Notes: The regression sample includes students who have 3rd and 8th grade math and reading test scores and 3rd grade student characteristics. All regressions control for state and year fixed effects, student race, gender, ethnicity, disability status, English language learner status, free or reduced-price lunch status, and enrollment status in special education.

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

Table 3: 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	0.793	1										
	3rd & 8th	0.846	0.994	1									
MA	3rd	0.997	0.792	0.844	1								
	8th	0.794	0.999	0.993	0.795	1							
	3rd & 8th	0.847	0.993	0.999	0.847	0.994	1						
NC	3rd	0.997	0.789	0.844	0.994	0.790	0.845	1					
	8th	0.790	0.999	0.993	0.788	0.998	0.992	0.789	1				
	3rd & 8th	0.844	0.993	0.999	0.841	0.991	0.998	0.844	0.994	1			
WA	3rd	0.996	0.790	0.843	0.996	0.794	0.846	0.994	0.787	0.841	1		
	8th	0.797	0.999	0.993	0.797	0.999	0.993	0.792	0.998	0.991	0.798	1	
	3rd & 8th	0.851	0.993	0.999	0.850	0.993	0.999	0.849	0.992	0.999	0.851	0.994	1

Table 4: Model Coefficients by State (3rd Grade)

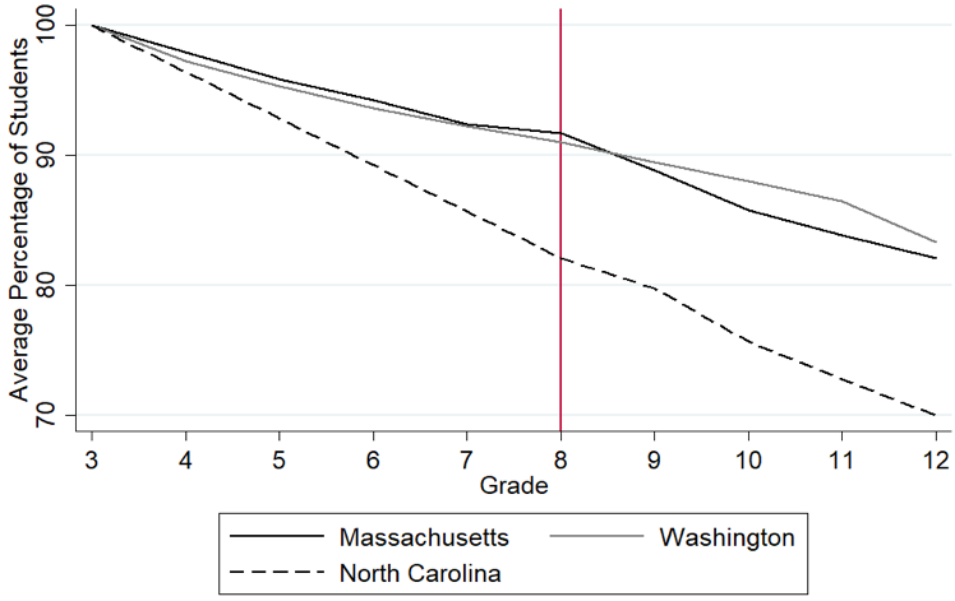
	High School Math Tests			Advanced Course-Taking			Graduation		
	MA	NC	WA	MA	NC	WA	MA	NC	WA
Math Percentile	0.485***	0.483***	0.541***	0.495 ***	0.483***	0.451***	0.109***	0.214***	0.142***
	(0.002)	(0.002)	(0.004)	(0.005)	(0.003)	(0.004)	(0.004)	(0.002)	(0.004)
Reading Percentile	0.179***	0.164***	0.171***	0.177***	0.243***	0.204***	0.061***	0.132***	0.113***
	(0.002)	(0.002)	(0.004)	(0.006)	(0.003)	(0.005)	(0.004)	(0.002)	(0.004)
R^2 or Pseudo-R^2	0.54	0.452	0.577	0.152	0.203	0.149	0.146	0.141	0.099
N	285,396	480,682	58,246	172,243	774,658	242,333	172,651	788,622	244,964

Notes: A The regression sample includes students who have 3rd grade math and reading test scores and 3rd grade student characteristics. All regressions control for year fixed effects, student race, gender, ethnicity, disability status, English language learner status, free or reduced-price lunch status, and enrollment status in special education.
** p<0.10* *** p<0.05* **** p<0.01.* Probability values are from a two-sided *t*-test.

Table 5: Correlations between Model Predictions by State

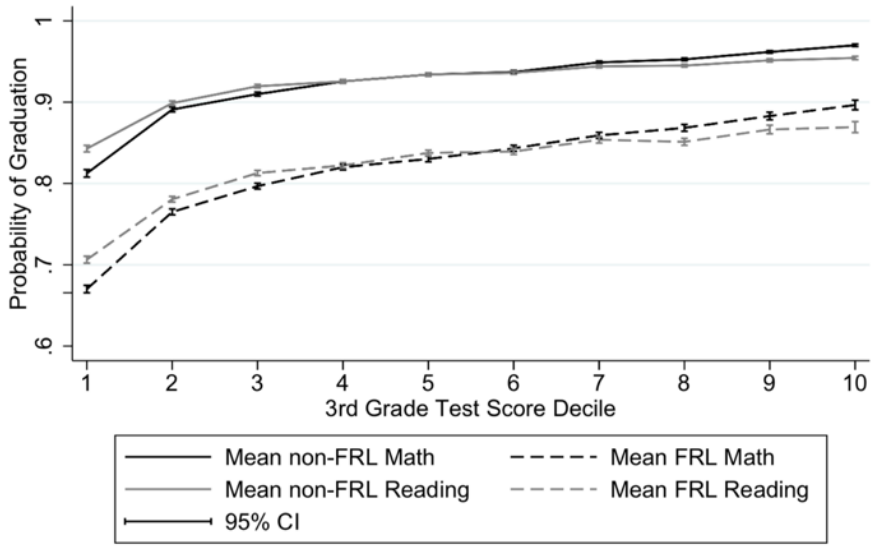
		MA Data			NC Data			WA Data		
		MA Model	NC Model	WA Model	MA Model	NC Model	WA Model	MA Model	NC Model	WA Model
Graduation	MA Model	1			1			1		
	NC Model	0.849	1		0.929	1		0.862	1	
	WA Model	0.938	0.891	1	0.958	0.835	1	0.915	0.892	1
Advanced Course-Taking	MA Model	1			1			1		
	NC Model	0.870	1		0.954	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



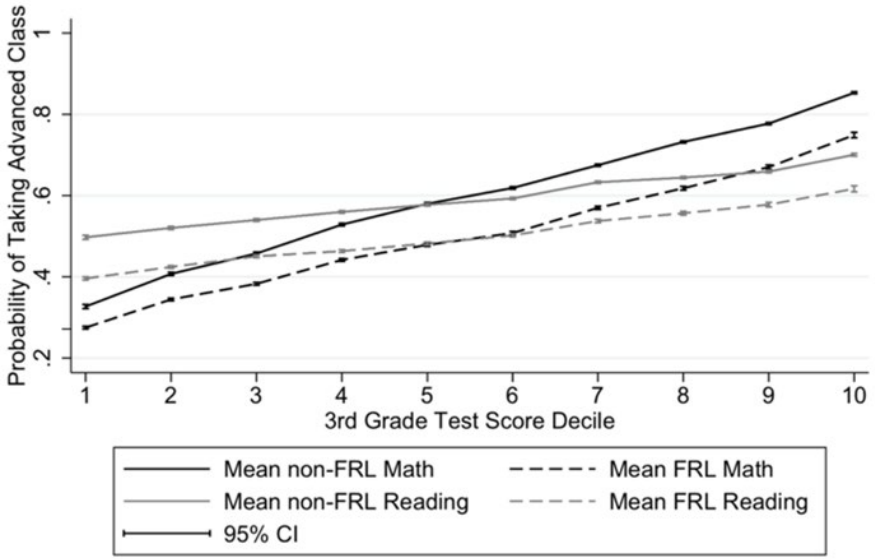
Notes: Average percent of observable 3rd grade students throughout the K-12 education system, broken up by state. Most student's observable in 8th grade have the "Above 50" outcome, and those observable through 12th grade have the "Graduation" outcome.

Figure 2: Probability of Graduation by 3rd Grade Test Score Decile and FRL



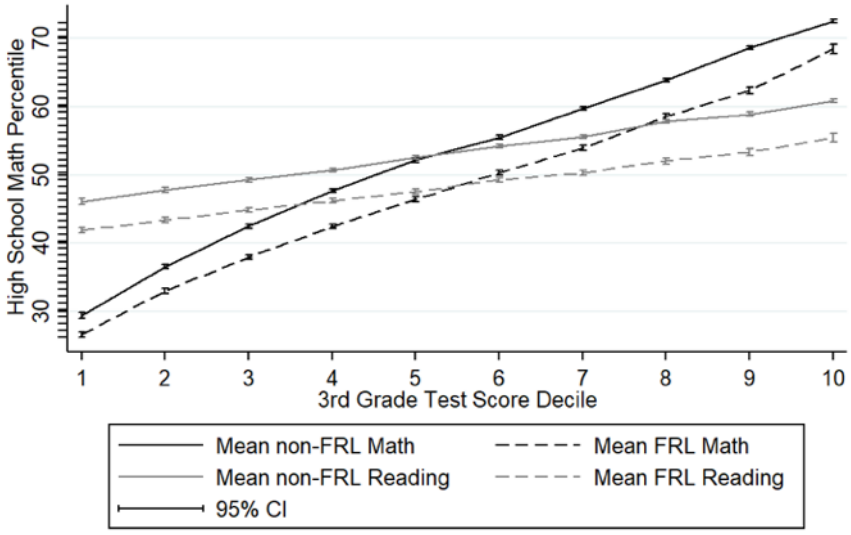
Notes: Probability of graduation by 3rd grade test score decile and FRL, estimated as marginal effects. Consistent, large effects of FRL status are seen, lowering probability of graduation by 8%-10% across all test scores for math and reading—approximately the same effect as going from the 1st to 10th decile of scores.

Figure 3: Probability of Advanced Course-Taking by 3rd Grade Test Score Decile and FRL



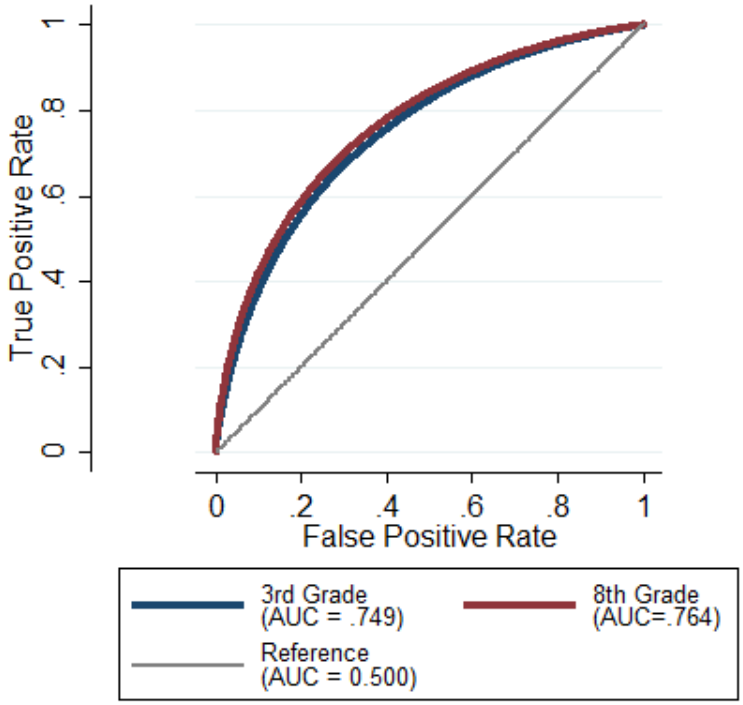
Notes: Probability of advanced course-taking by 3rd grade test score decile and FRL, estimated as marginal effects. Relatively consistent, large effects of FRL status are seen, lowering probability of advanced course-taking by 8%-10% across all test scores for math and reading—approximately the same effect as improving test scores by one decile.

Figure 4: High School Math Percentile by 3rd Grade Test Scores and FRL



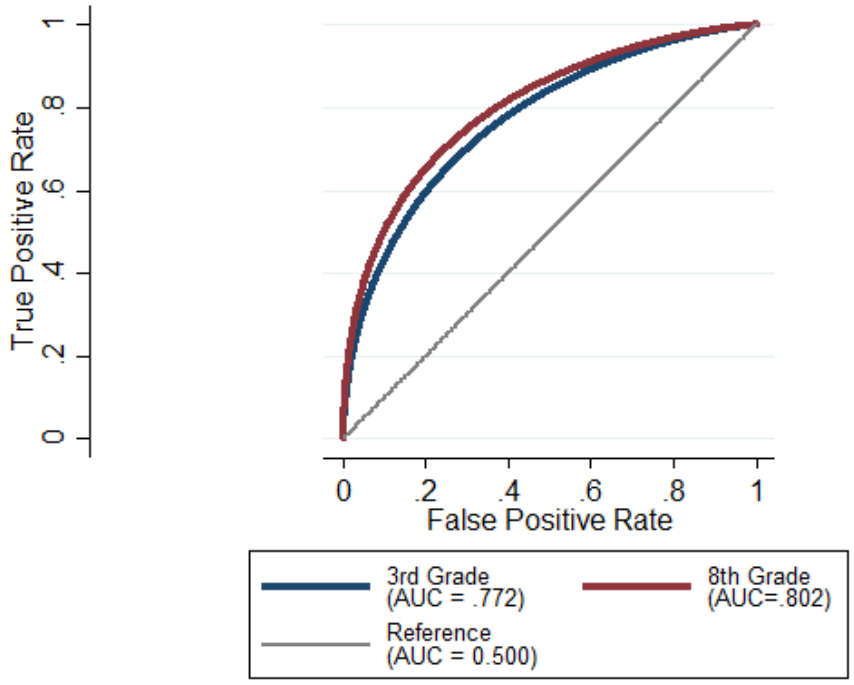
Notes: High school math percentile by 3rd grade test score decile and FRL, estimated as marginal effects. Relatively consistent, large effects of FRL status are seen, lowering high school math percentile by 3%-5% across all test scores for math and reading—a slightly smaller effect as improving test scores by one decile.

Figure 5: ROC Curve Predicting Graduation using 3rd and 8th Grade Test Scores



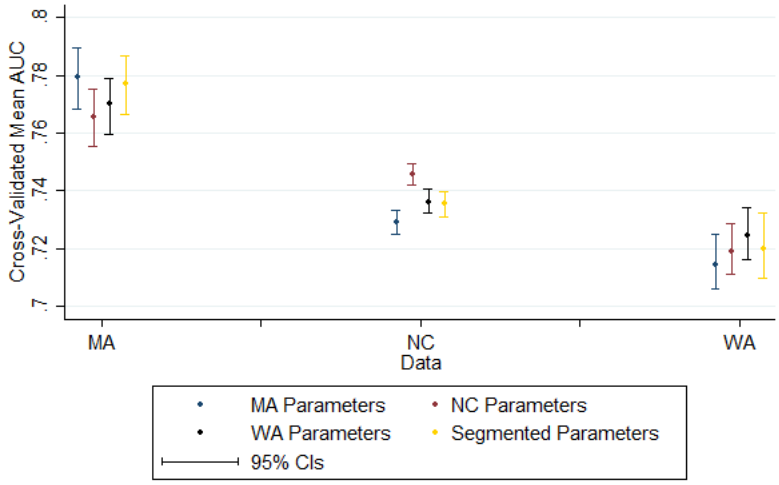
Notes: ROC curves corresponding to graduation prediction using both 3rd grade test scores and 8th grade test scores, with reported AUCs in the legend. The similarity of AUC and general shape of ROC curve shows a strong capacity for effective intervention targeting early in students' academic careers—as early as 3rd grade.

Figure 6: ROC Curve Predicting Advanced Course-Taking using 3rd and 8th Grade Test Scores



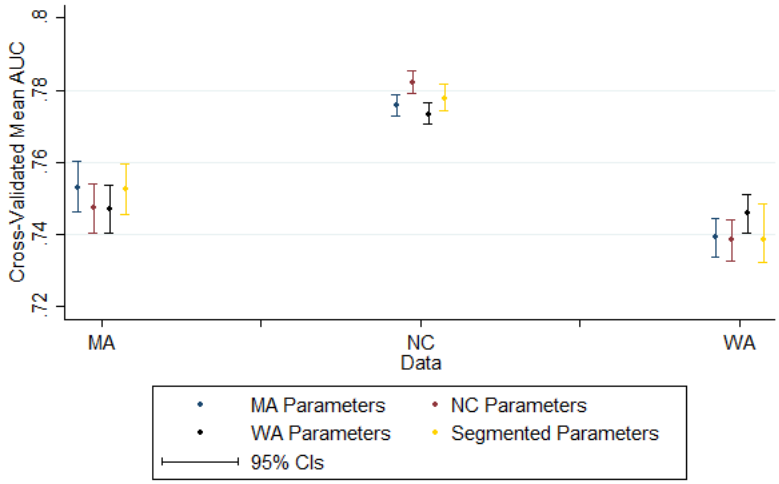
Notes: ROC curves corresponding to advanced course-taking prediction using both 3rd grade test scores and 8th grade test scores, with reported AUCs in the legend. The similarity of AUC and general shape of ROC curve shows a strong capacity for predicting high achievement as early as 3rd grade.

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



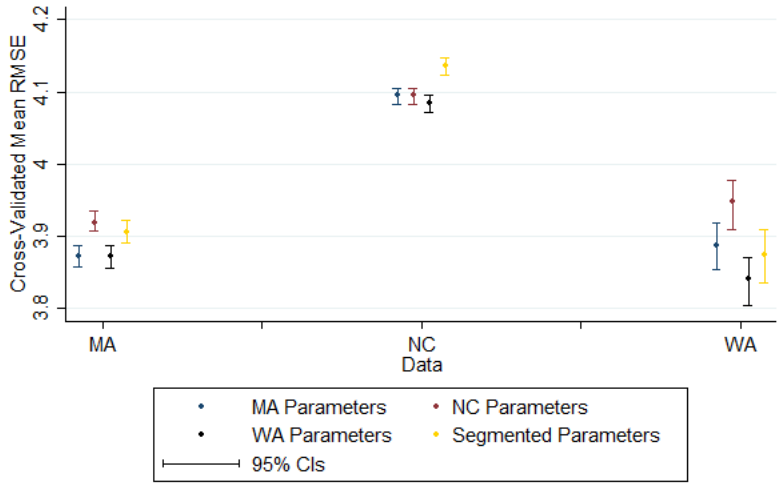
Notes: Mean estimated probabilities of 10-fold cross-validated AUC for graduation. Confidence intervals are generated by repeating 10-fold CV over 100 iterations.

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



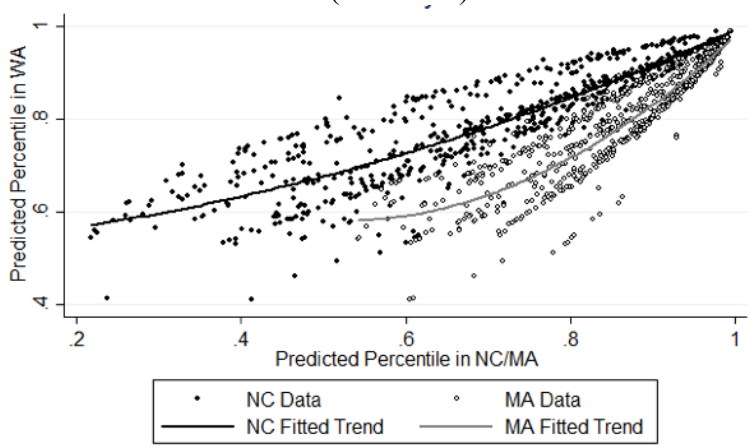
Notes: Mean estimated probabilities of 10-fold cross-validated AUC for advanced course-taking. Confidence intervals are generated by repeating 10-fold CV over 100 iterations.

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



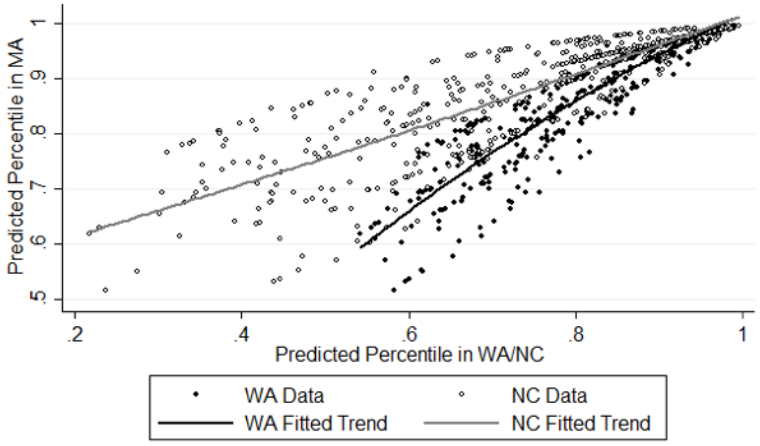
Notes: Mean estimates of 10-fold cross-validated RMSE for high school math tests. Confidence intervals are generated by repeating 10-fold CV over 100 iterations.

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



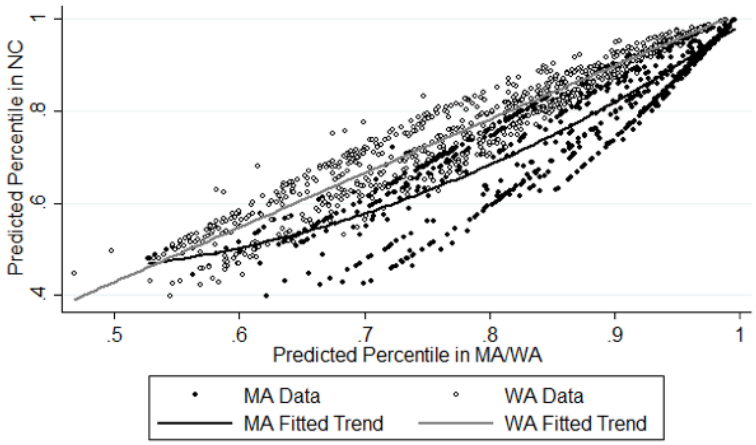
Notes: Scatterplot of predicted probabilities of graduation in Washington compared to predicted probabilities in North Carolina and Massachusetts, estimated on students in Washington State. Points displayed are a random subset of < 5% of the data, where the probability of displaying a point is inversely 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)



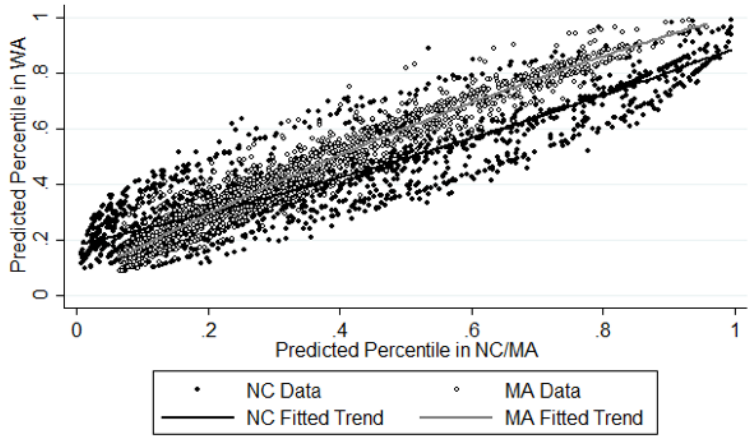
Notes: Scatterplot of predicted probabilities of graduation in Massachusetts compared to predicted probabilities in North Carolina and Washington, estimated on students in Massachusetts. Points displayed are a random subset of < 5% of the data, where the probability of displaying a point is inversely 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)



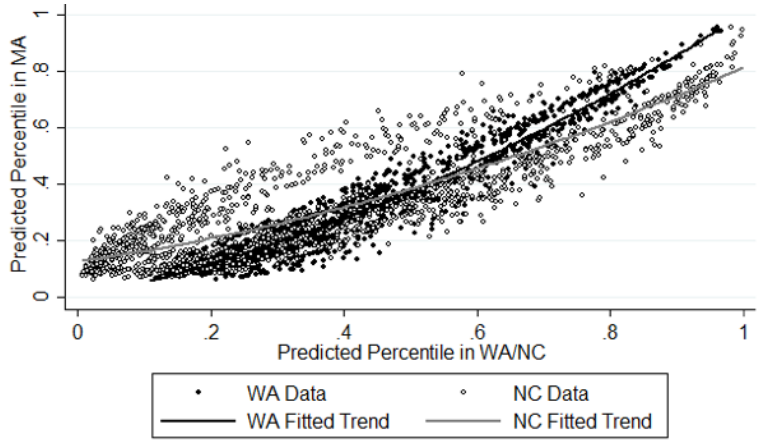
Notes: Scatterplot of predicted probabilities of graduation in North Carolina compared to predicted probabilities in Massachusetts and Washington, estimated on students in North Carolina. Points displayed are a random subset of < 5% of the data, where the probability of displaying a point is inversely 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)



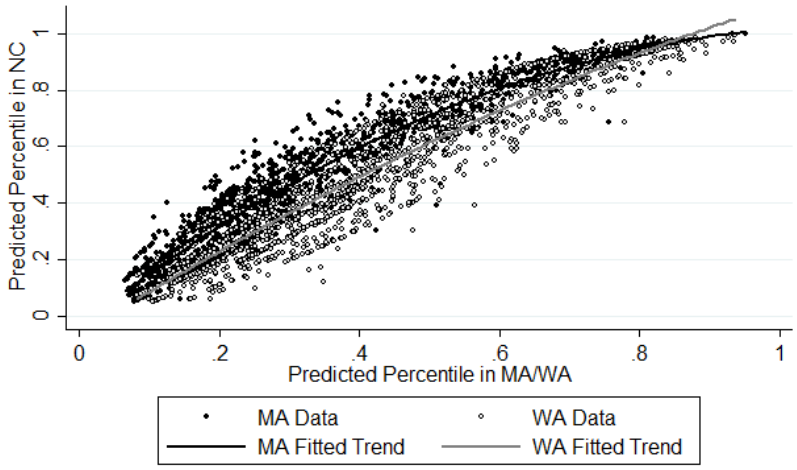
Notes: Scatterplot of predicted probabilities of advanced course-taking in Washington compared to predicted probabilities in Massachusetts and North Carolina, estimated on students in Washington.

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



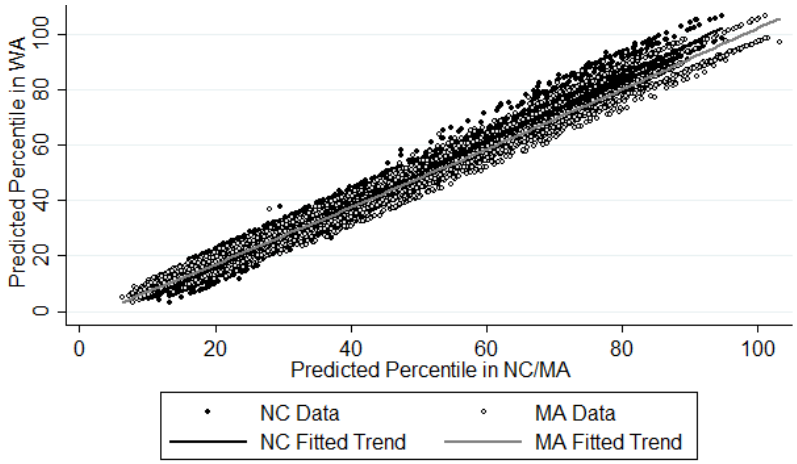
Notes: Scatterplot of predicted probabilities of advanced course-taking in Massachusetts compared to predicted probabilities in Washington and North Carolina, estimated on students in Massachusetts.

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



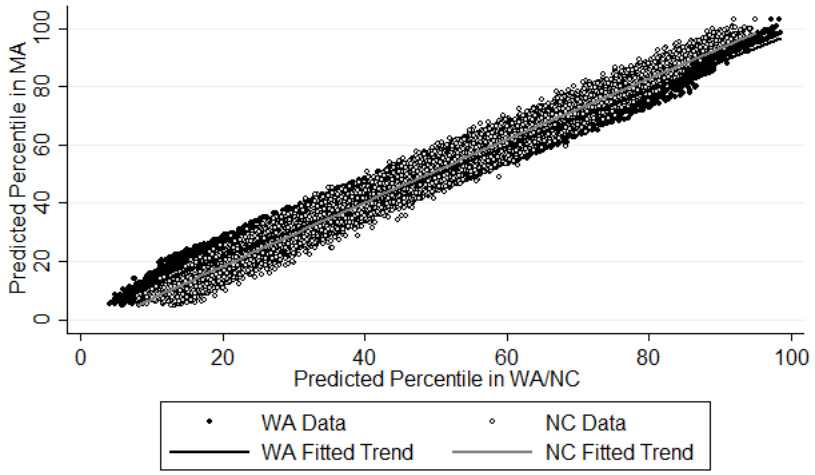
Notes: Scatterplot of predicted probabilities of advanced course-taking in North Carolina compared to predicted probabilities in Washington and Massachusetts, estimated on students in North Carolina.

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



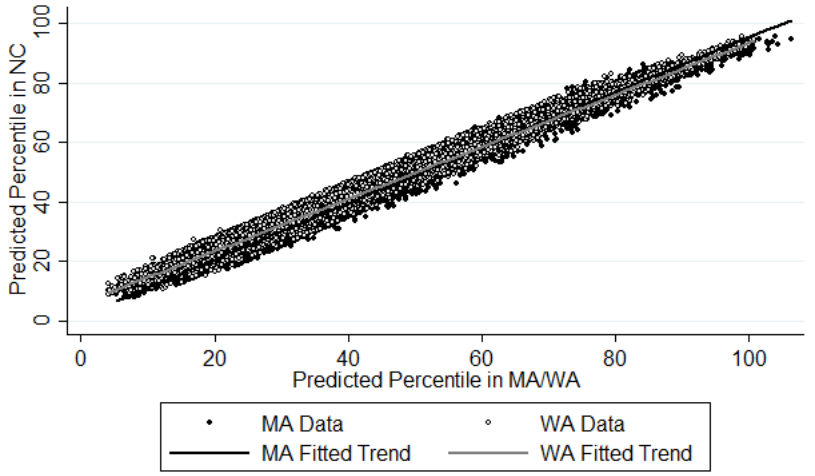
Notes: Scatterplot of predicted percentiles of high school math test in Washington compared to predicted probabilities in North Carolina and Massachusetts, estimated on students in Washington.

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



Notes: 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.

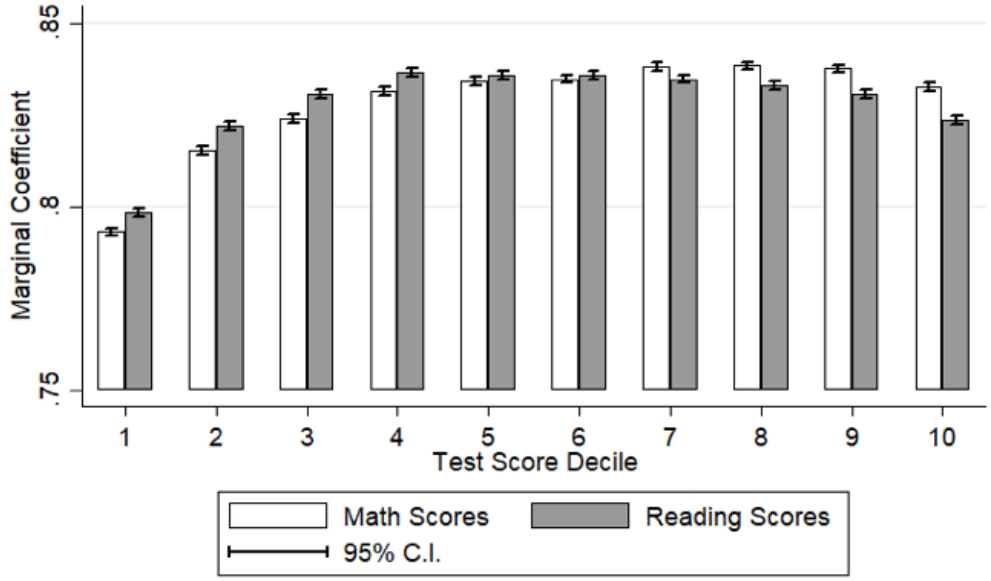
Figure 18: Scatterplot of Predicted High School Math Percentile in NC vs Predicted Probabilities from Out-of-State Models (3rd Grade)



Notes: 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.

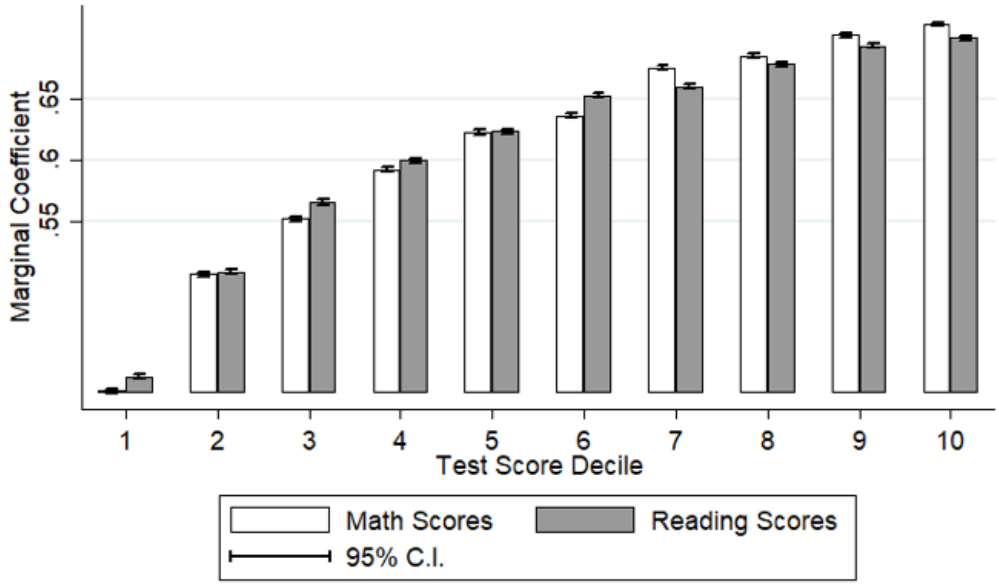
Appendix A

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



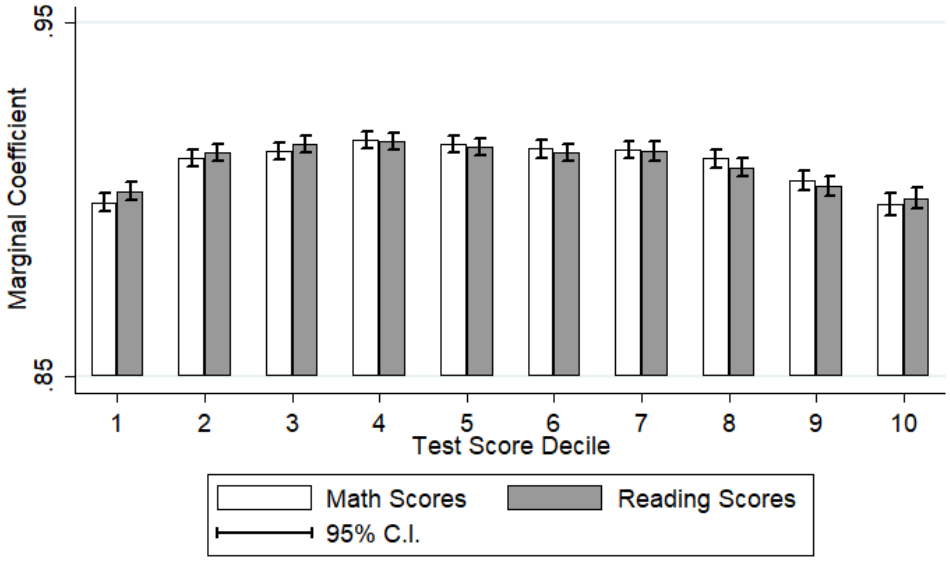
Notes: Marginal probabilities of sample persistence by 8th grade, controlling for student effects, broken up by test score decile. Students in the lowest decile of test scores are significantly less likely to persist in the sample through 8th grade, and students in the highest decile of test scores are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

Figure A2: Marginal Effects of Test Score on Sample Persistence by 12th Grade



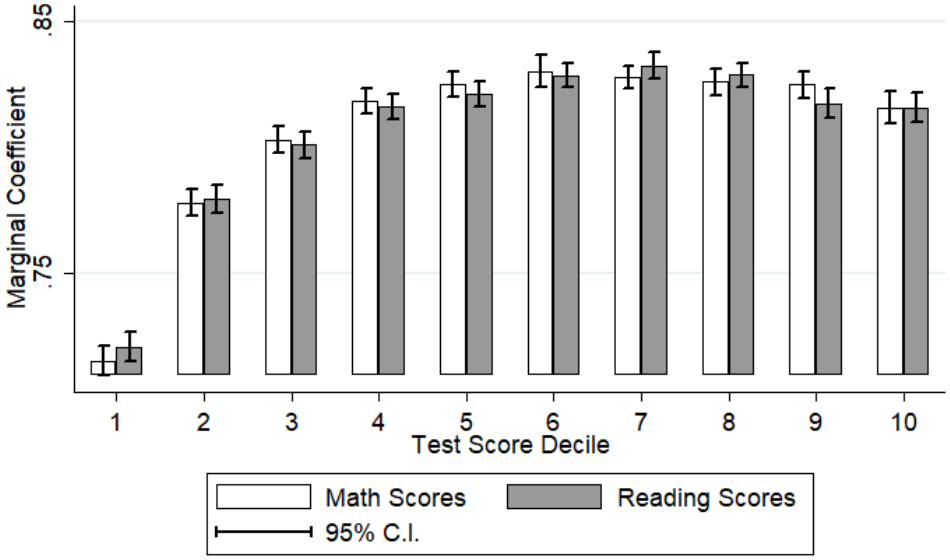
Notes: Marginal probabilities of sample persistence by 12th grade, controlling for student effects, broken up by test score decile. Students in the lowest decile of test scores are significantly less likely to persist in the sample through 12th grade.

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



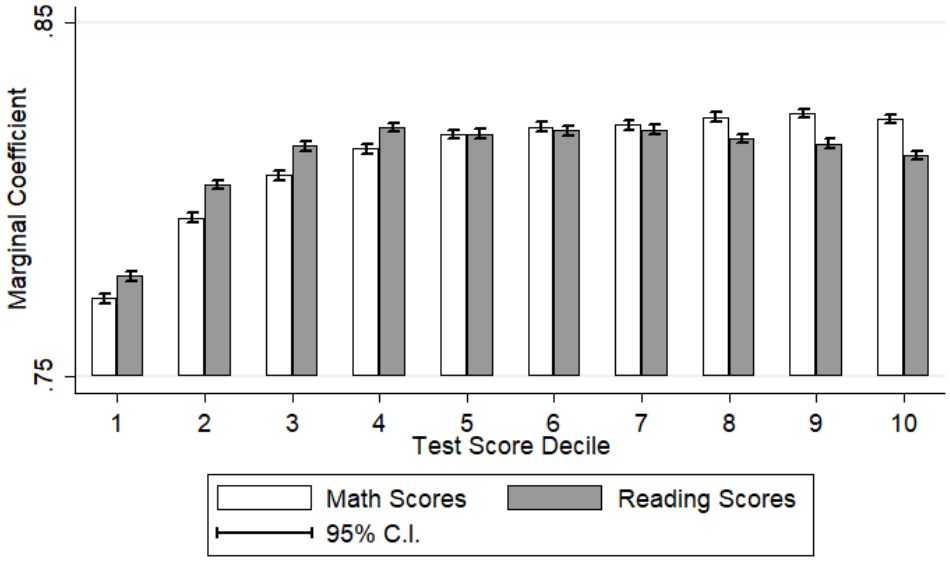
Notes: Marginal probabilities of sample persistence by 8th grade in MA, controlling for student effects, broken up by test score decile. Students in the lowest and highest decile of test scores are significantly less likely to persist in the sample through 8th grade.

Figure A4: Marginal Effects of Test Score on Sample Persistence by 12th Grade (MA)



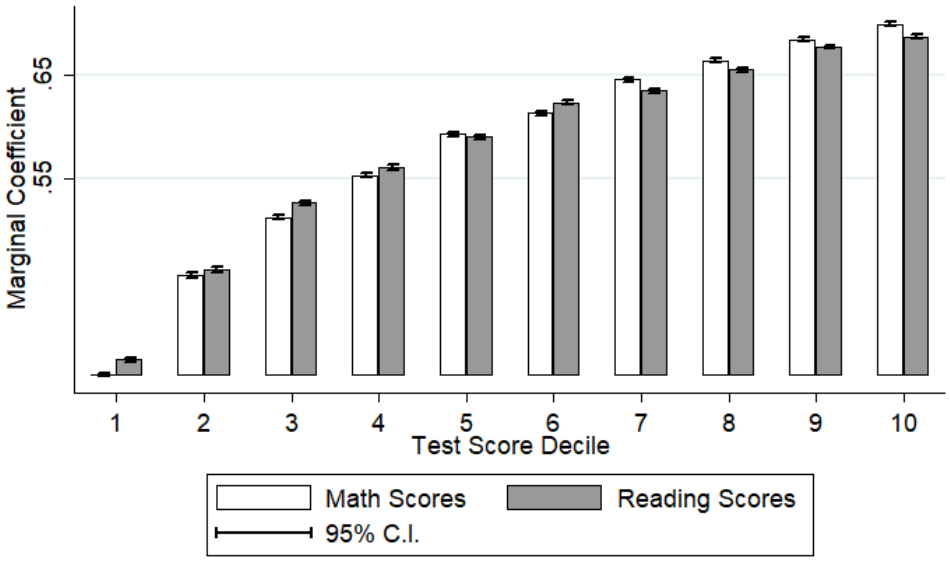
Notes: Marginal probabilities of sample persistence by 12th grade in MA, controlling for student effects, broken up by test score decile. Students in the highest decile of test scores are somewhat less likely to persist in the sample through 8th grade, but students in the lowest decile of scores are significantly less likely to persist.

Figure A5: Marginal Effects of Test Score on Sample Persistence by 8th Grade (NC)



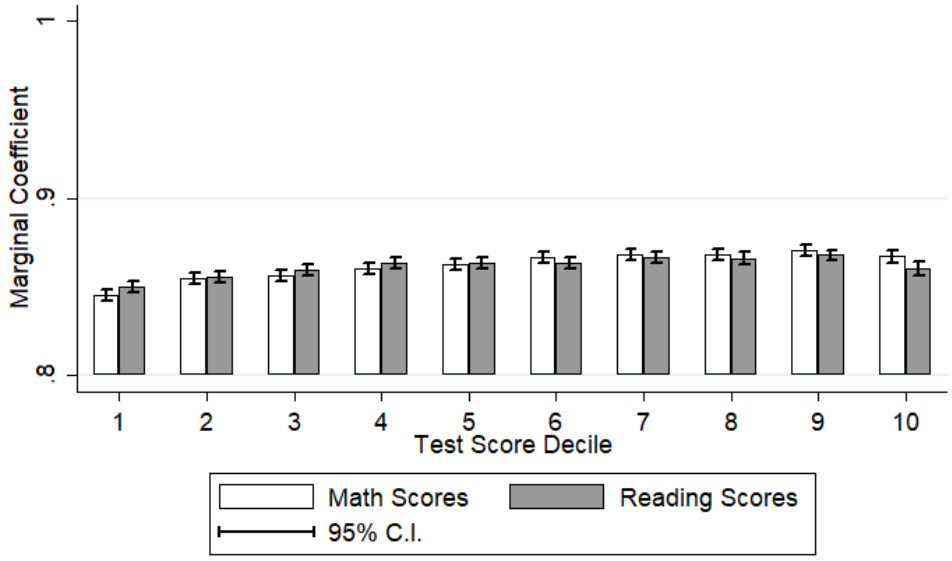
Notes: Marginal probabilities of sample persistence by 8th grade in NC, controlling for student effects, broken up by test score decile. Students in the lowest decile of test scores are significantly less likely to persist in the sample through 8th grade.

Figure A6: Marginal Effects of Test Score on Sample Persistence by 12th Grade (NC)



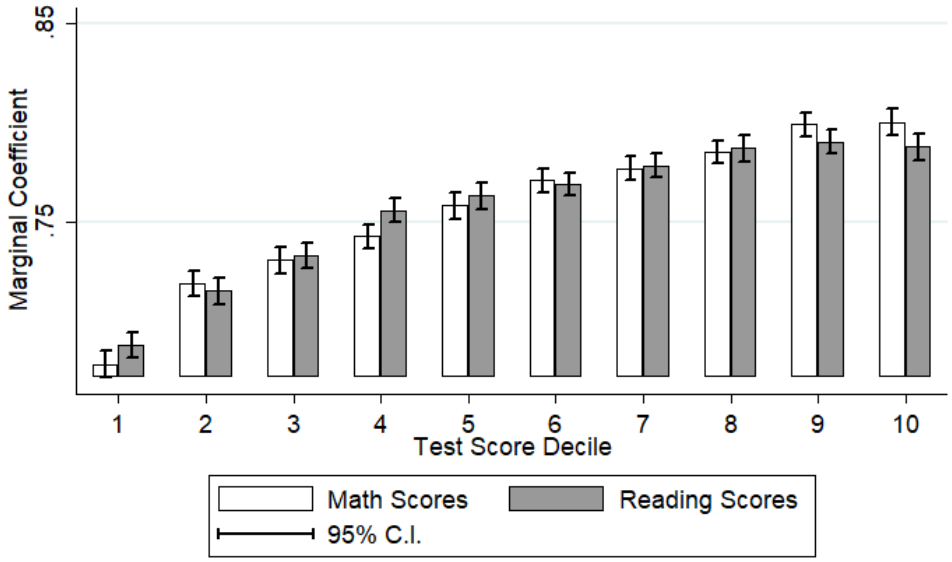
Notes: Marginal probabilities of sample persistence by 12th grade in NC, controlling for student effects, broken up by test score decile. Students in the lowest decile of test scores are significantly less likely to persist in the sample through 12th grade.

Figure A7: 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.

Figure A8: Marginal Effects of Test Score on Sample Persistence by 12th Grade (WA)



Notes: Marginal probabilities of sample persistence by 12th grade in WA, 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 12th grade.

Figure A9: ROC Curve of Graduation by Model Specification in Massachusetts

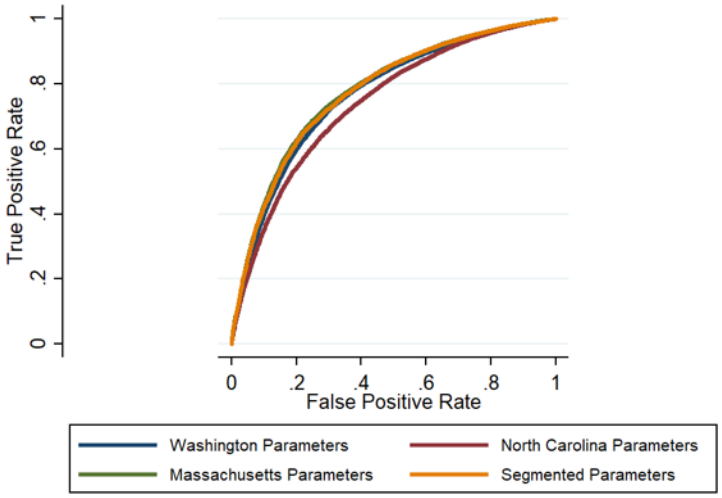


Figure A10: ROC Curve of Graduation by Model Specification in Washington

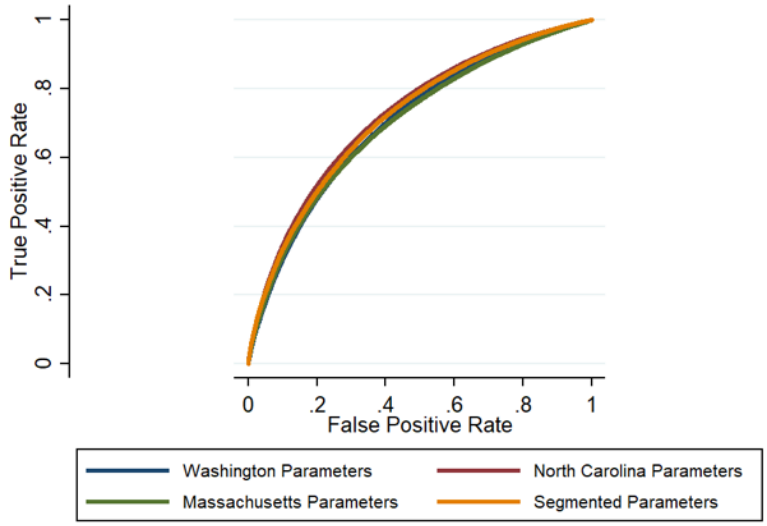


Figure A11: ROC Curve of Graduation by Model Specification in North Carolina

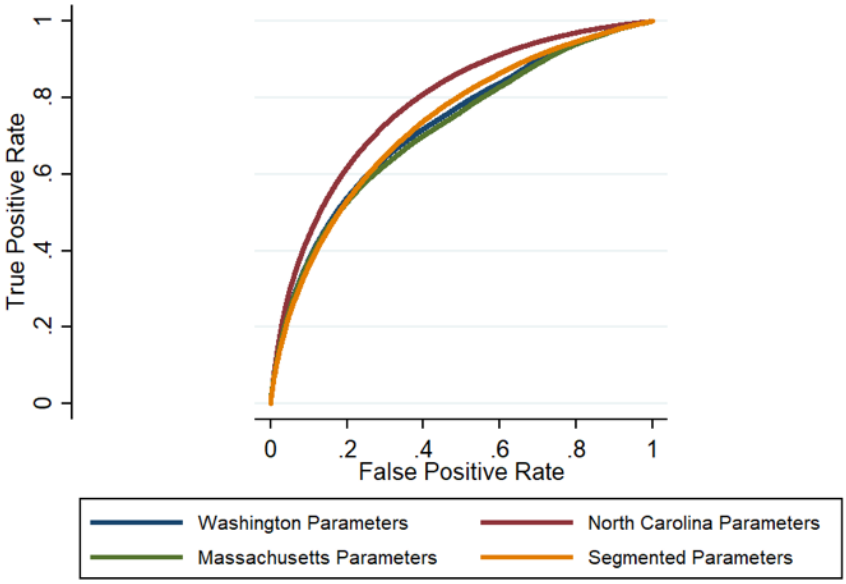


Figure A12: Probability of Graduation by 3rd Grade Test Score Decile and FRL in Washington

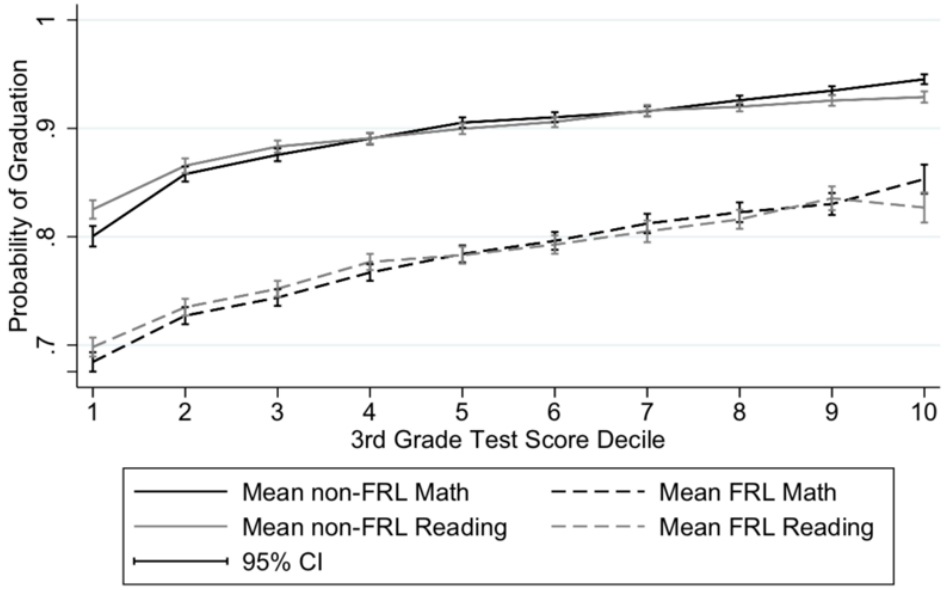


Figure A13: Probability of Graduation by 3rd Grade Test Score Decile and FRL in Massachusetts

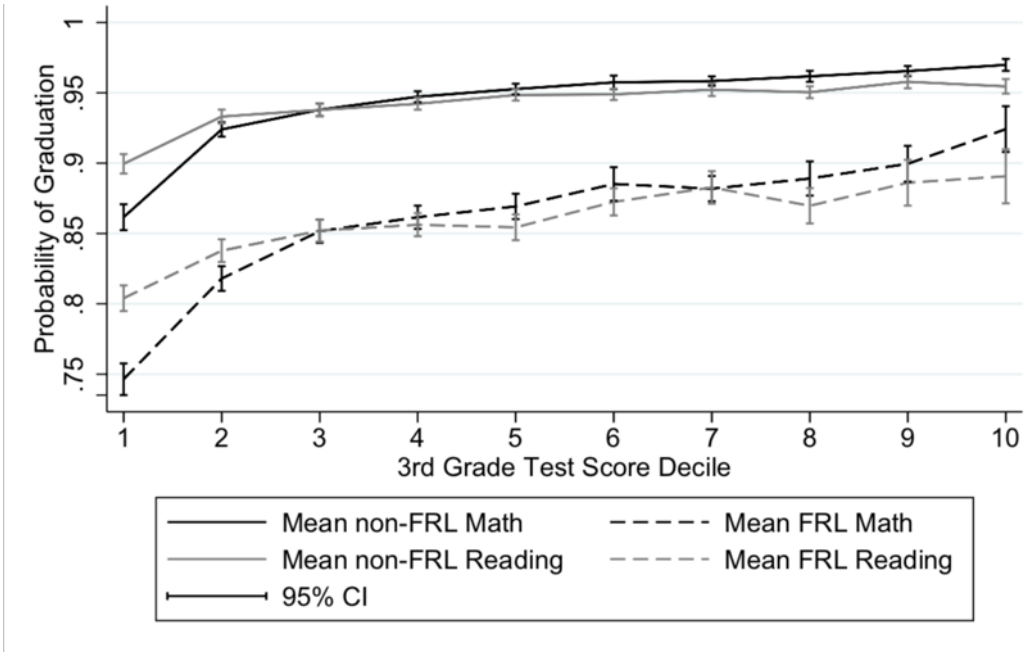


Figure A14: Probability of Graduation by 3rd Grade Test Score Decile and FRL in North Carolina

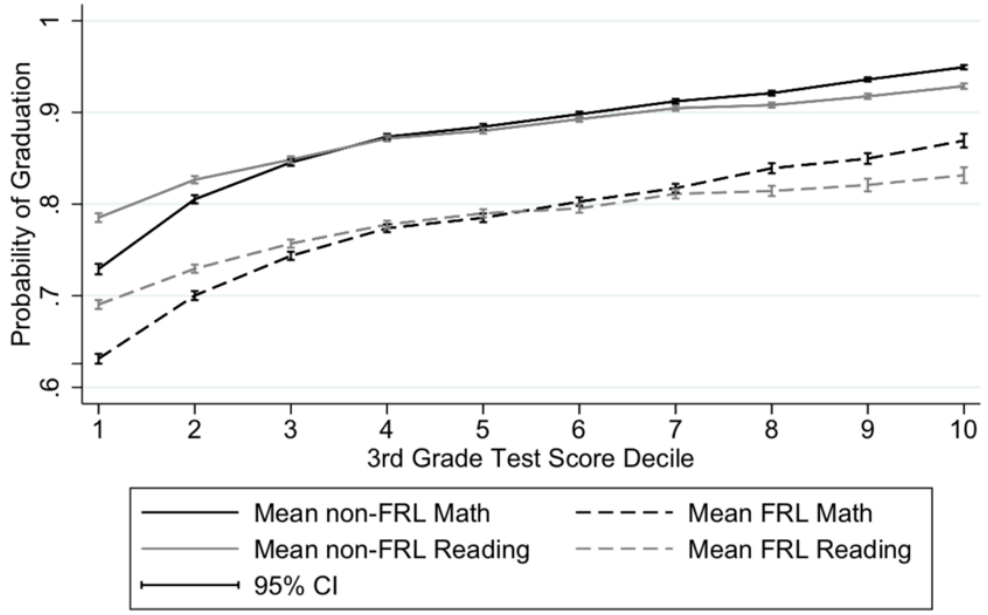


Figure A15: Probability of Advanced Course-Taking by 3rd Grade Test Score Decile and FRLin Washington

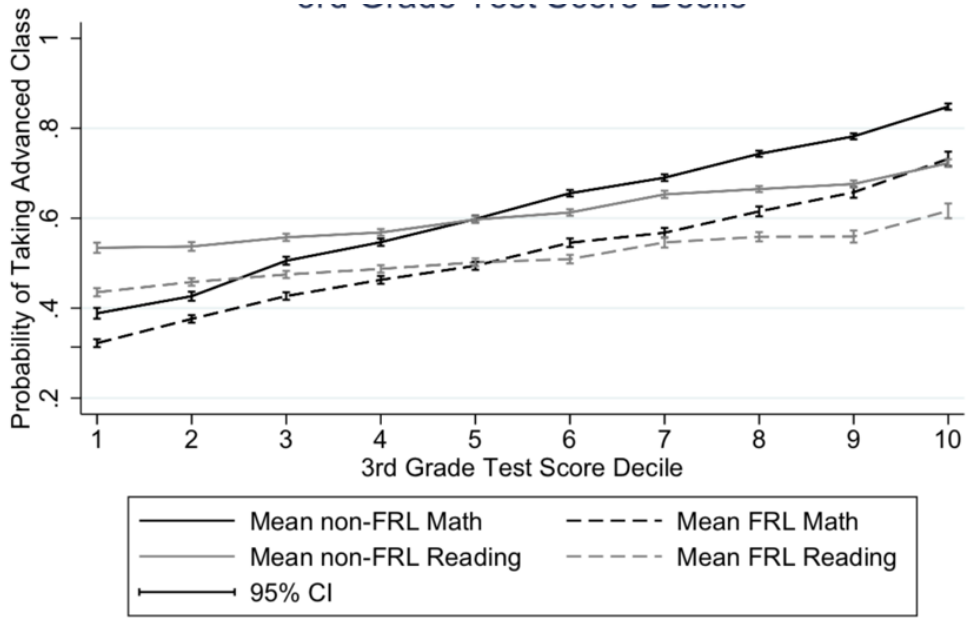


Figure A16: Probability of Advanced Course-Taking by 3rd Grade Test Score Decile and FRLin Massachusetts

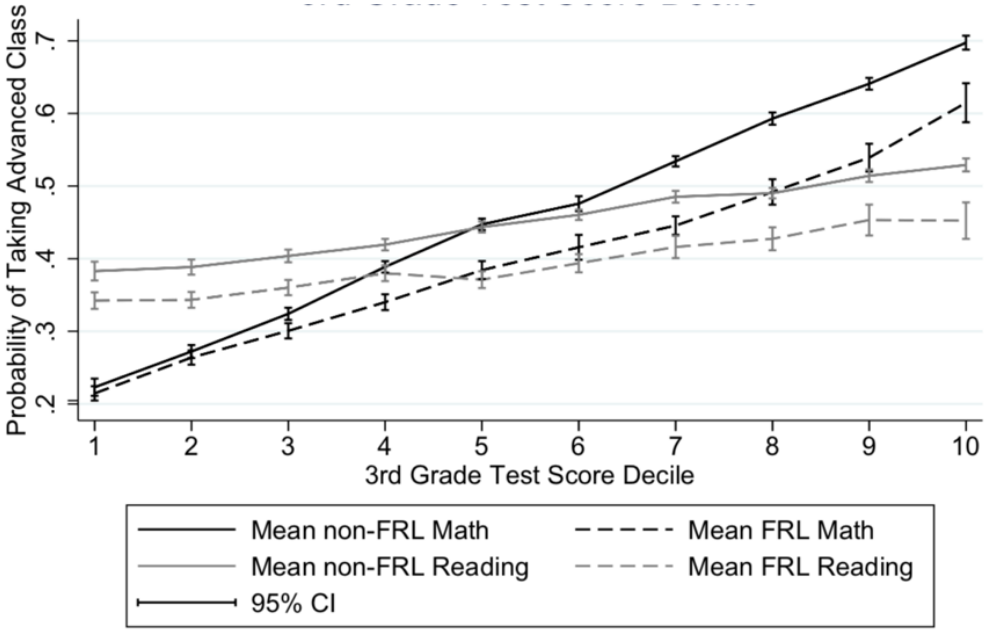


Figure A17: Probability of Advanced Course-Taking by 3rd Grade Test Score Decile and FRLin North Carolina

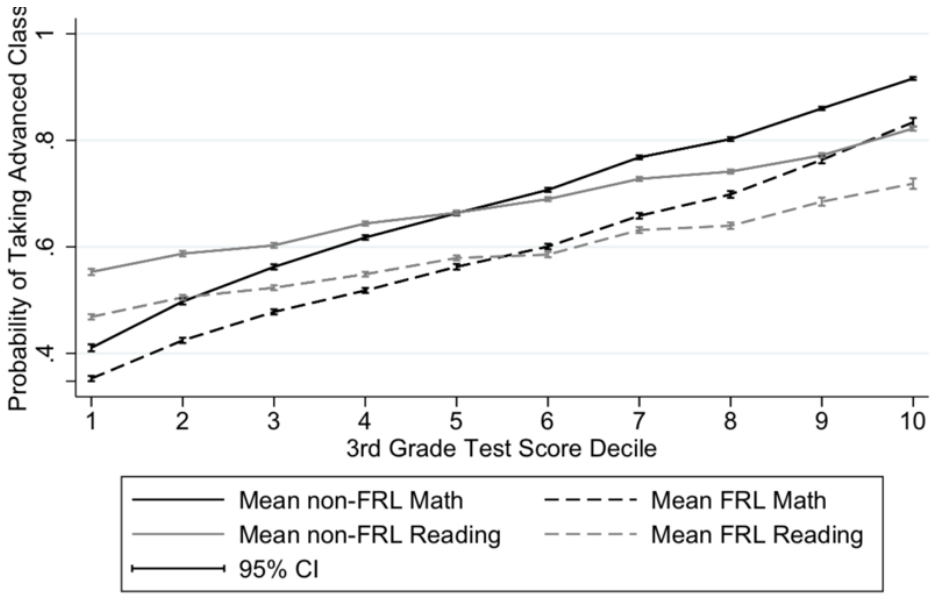


Figure A18: High School Math Percentile by 3rd Grade Test Score Decile and FRLin Washington

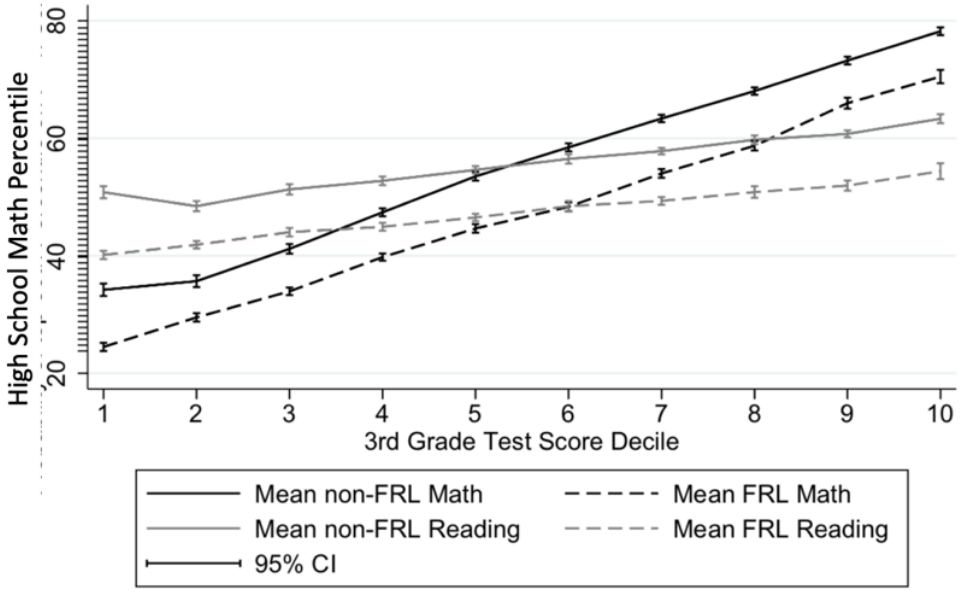
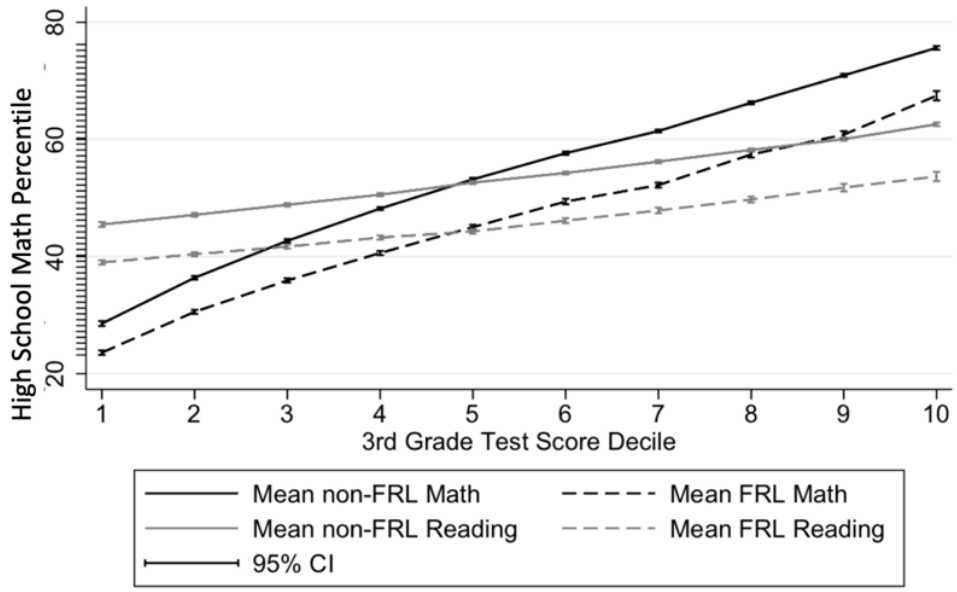
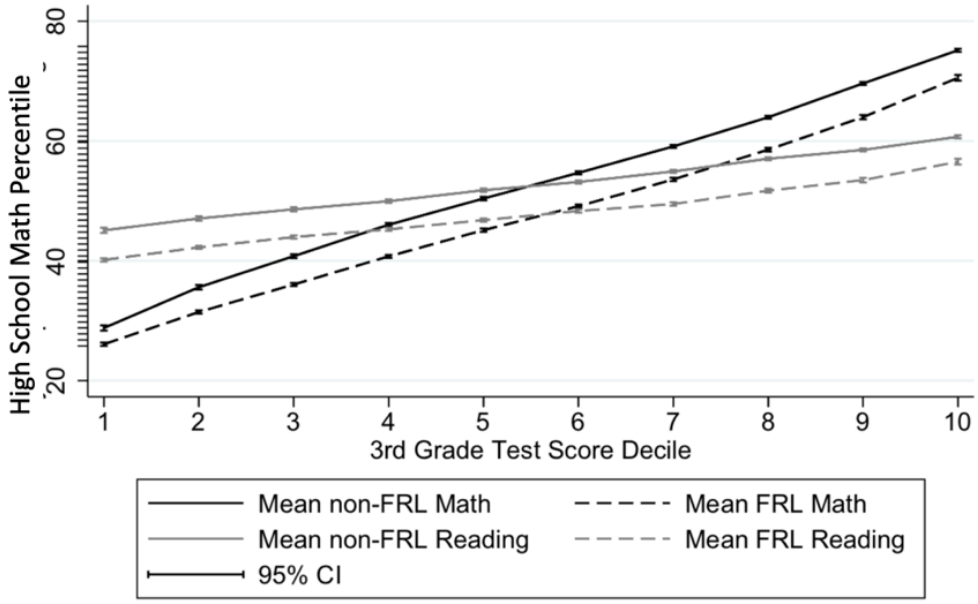


Figure A19: High School Math Percentile by 3rd Grade Test Score Decile and FRLin Massachusetts



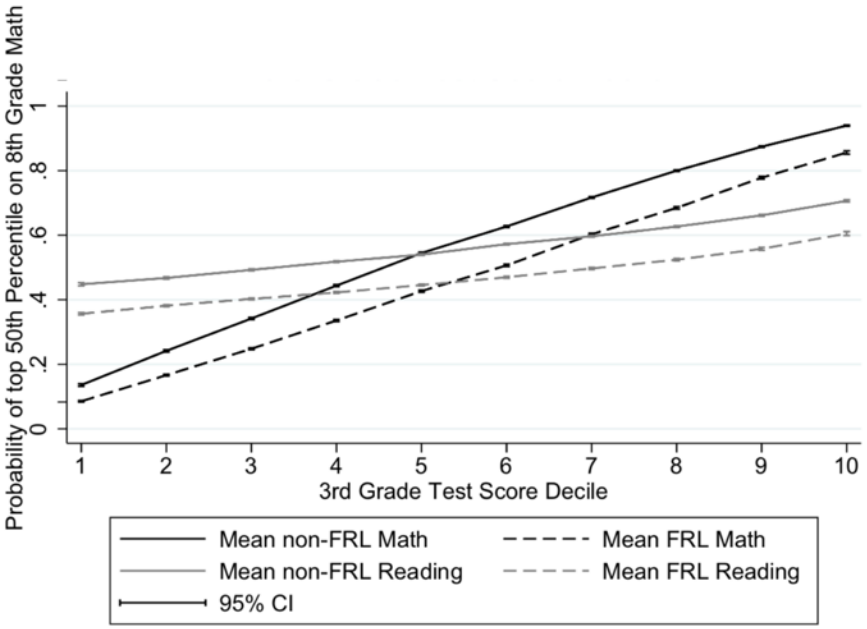
Notes: Probability of top 50th percentile in 8th grade math by 3rd grade test score decile and FRL, estimated as marginal effects. Large effects of FRL status are seen, lowering probability of top-half achievement by up to 10% for math and reading—approximately the same effect as a one-decile change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

Figure A20: High School Math Percentile by 3rd Grade Test Score Decile and FRLin North Carolina



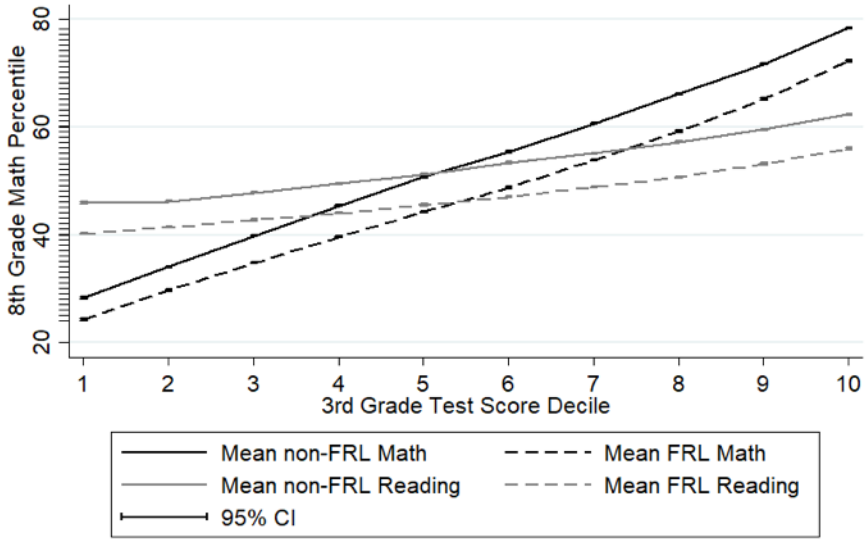
Notes: Probability of top 50th percentile in 8th grade math by 3rd grade test score decile and FRL, estimated as marginal effects. Large effects of FRL status are seen, lowering probability of top-half achievement by up to 10% for math and reading—approximately the same effect as a one-decile change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

Figure A21: Probability of Top 50th Percentile in 8th Grade Math by 3rd Grade Test Scores and FRL



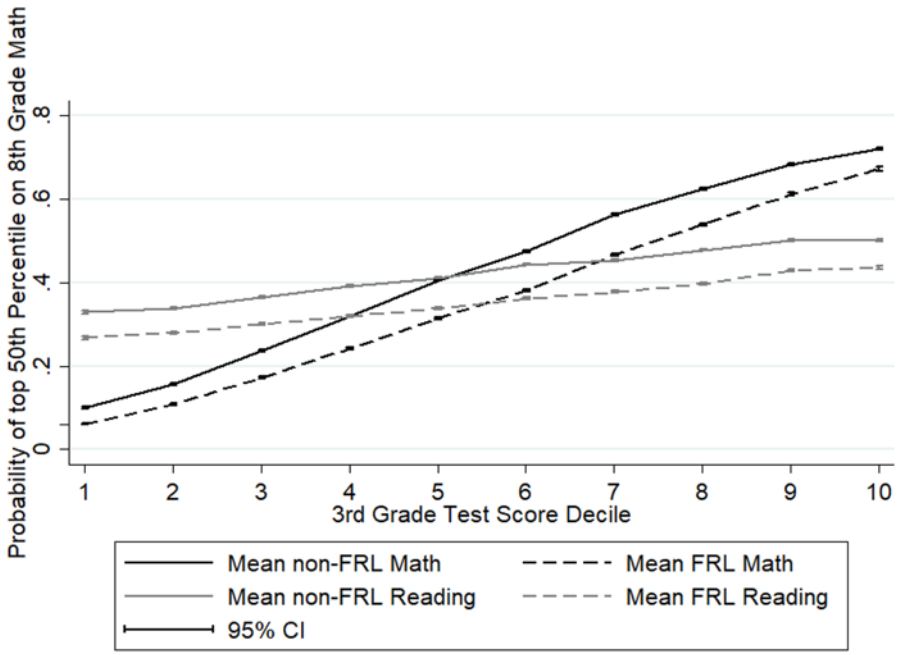
Notes: Probability of top 50th percentile in 8th grade math by 3rd grade test score decile and FRL, estimated as marginal effects. Large effects of FRL status are seen, lowering probability of top-half achievement by up to 10% for math and reading—approximately the same effect as a one-decile change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

Figure A22: 8th Grade Math Percentile by 3rd Grade Test Scores and FRL



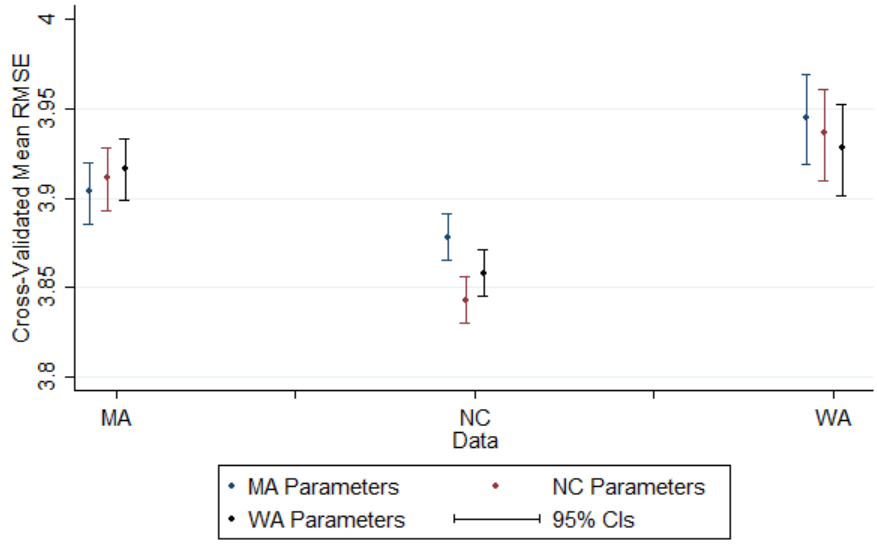
Notes: 8th grade math percentile by 3rd grade test score decile and FRL, estimated as marginal effects. Large effects of FRL status are seen, lowering percentile by up to 10 for math and reading.

Figure A23: Probability of Top 50th Percentile in High School Math by 3rd Grade Test Scores and FRL



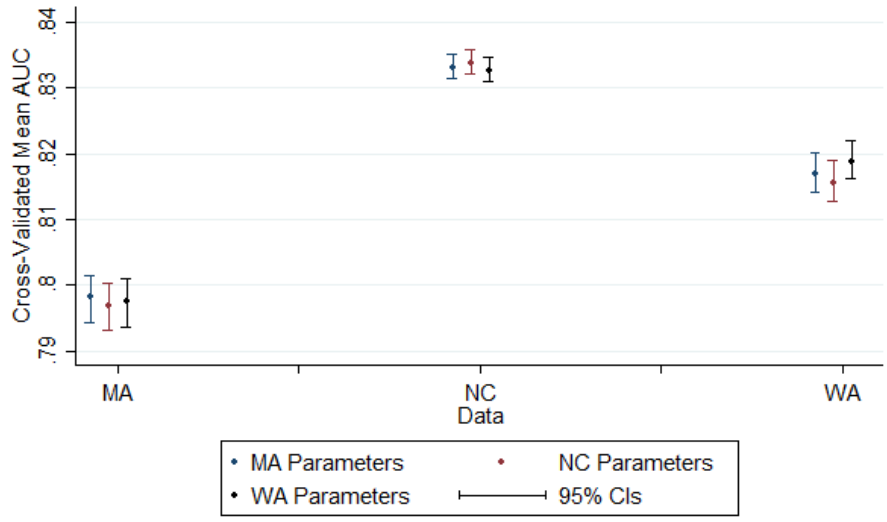
Notes: 8th grade math percentile by 3rd grade test score decile and FRL, estimated as marginal effects. Large effects of FRL status are seen, lowering percentile by up to 10 for math and reading.

Figure A24: 8th Grade Math Percentile Cross-Validated RMSE Estimates by Prediction Model



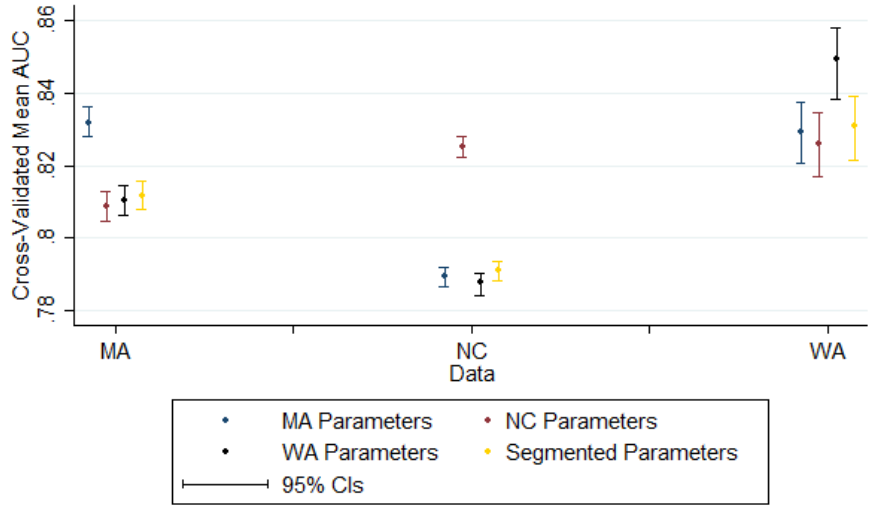
Notes: Mean estimates of 10-fold cross-validated RMSE for 8th grade math tests. Confidence intervals are generated by repeating 10-fold CV over 100 iterations.

Figure A25: Probability of Top 50th Percentile in 8th Grade Math Cross-Validated AUC Estimated by Prediction Model



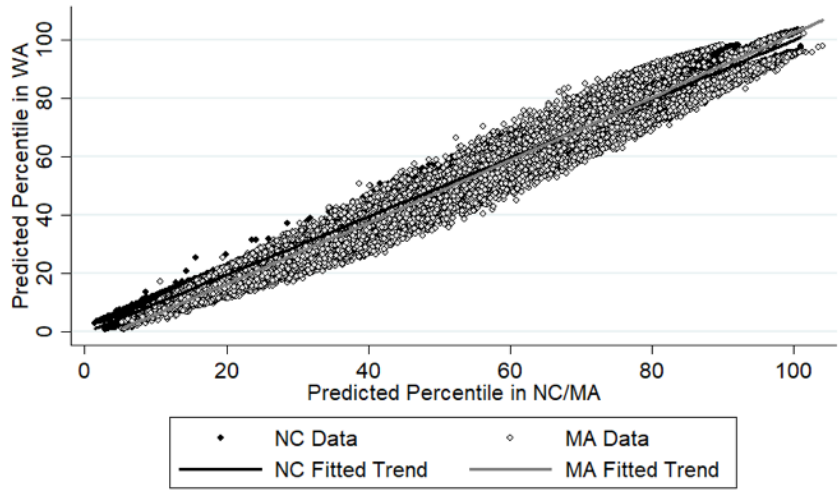
Notes: Mean estimates of 10-fold cross-validated AUC for the probability of scoring in the top half of 8th grade math test scores. Confidence intervals are generated by repeating 10-fold CV over 100 iterations

Figure A26: Probability of Top 50th Percentile High School Math Cross-Validated AUC Estimated by Prediction Model



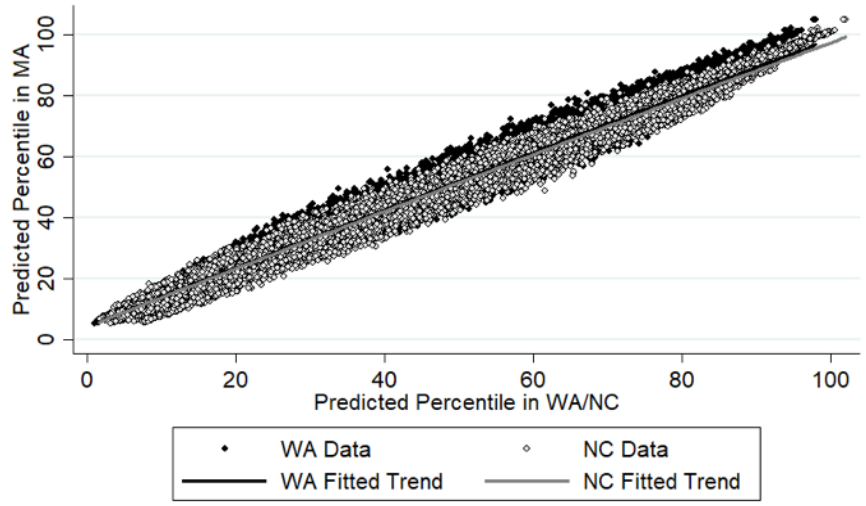
Notes: Mean estimates of 10-fold cross-validated AUC for the probability of scoring in the top half of high school math test scores. Confidence intervals are generated by repeating 10-fold CV over 100 iterations.

Figure A27: Scatterplot of Predicted 8th Grade Math Percentile in WA vs Predicted Probabilities from Out-of-State Models (3rd Grade)



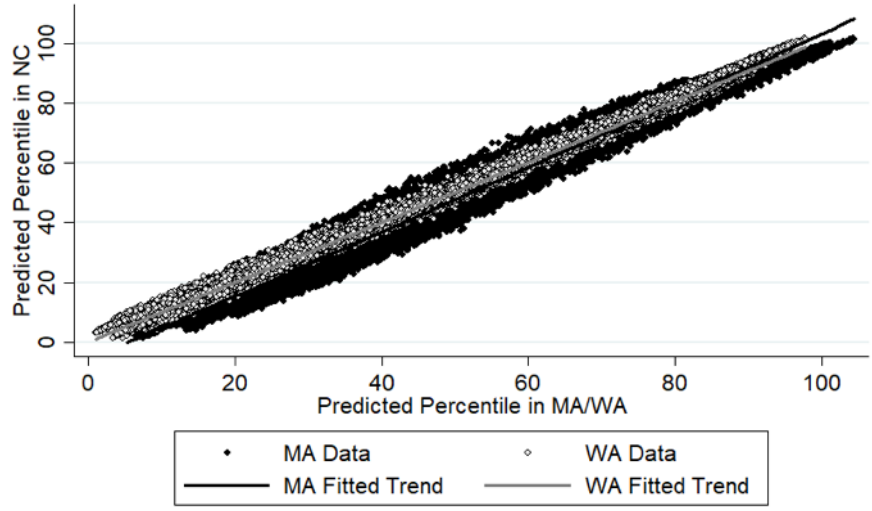
Notes: Scatterplot of predicted percentiles of 8th grade math test in Washington compared to predicted probabilities in North Carolina and Massachusetts, estimated on students in Washington.

Figure A28: Scatterplot of Predicted 8th Grade Math Percentile in MA vs Predicted Probabilities from Out-of-State Models (3rd Grade)



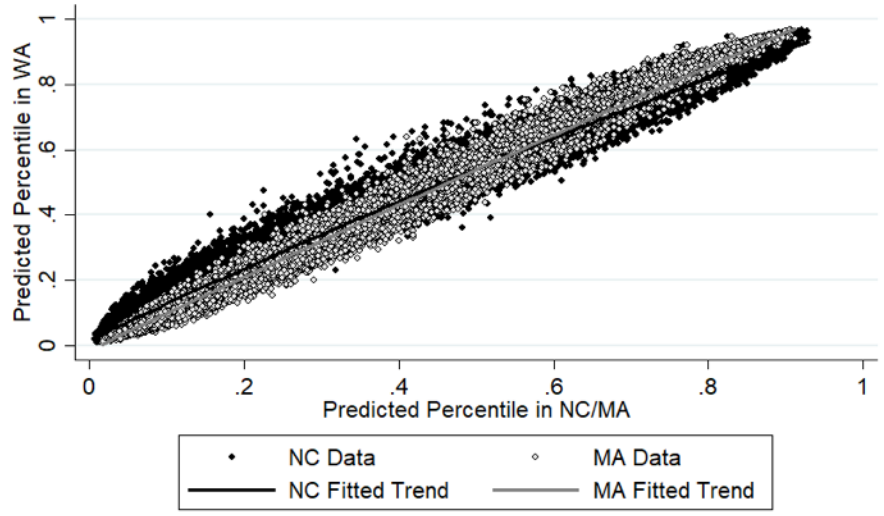
Notes: Scatterplot of predicted percentiles of 8th grade math test in Massachusetts compared to predicted probabilities in North Carolina and Washington, estimated on students in Massachusetts.

Figure A29: Scatterplot of Predicted 8th Grade Math Percentile in NC vs Predicted Probabilities from Out-of-State Models (3rd Grade)



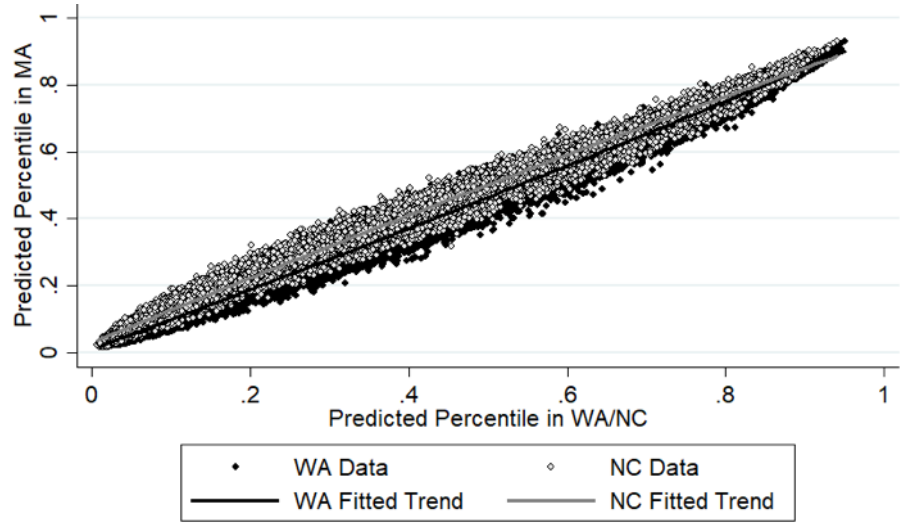
Notes: Scatterplot of predicted percentiles of 8th grade math test in North Carolina compared to predicted probabilities in Massachusetts and Washington, estimated on students in North Carolina.

Figure A30: Scatterplot of Predicted Probabilities of Top Half 8th Grade Math Tests in WA vs Predicted Probabilities from Out-of-State Models (3rd Grade)



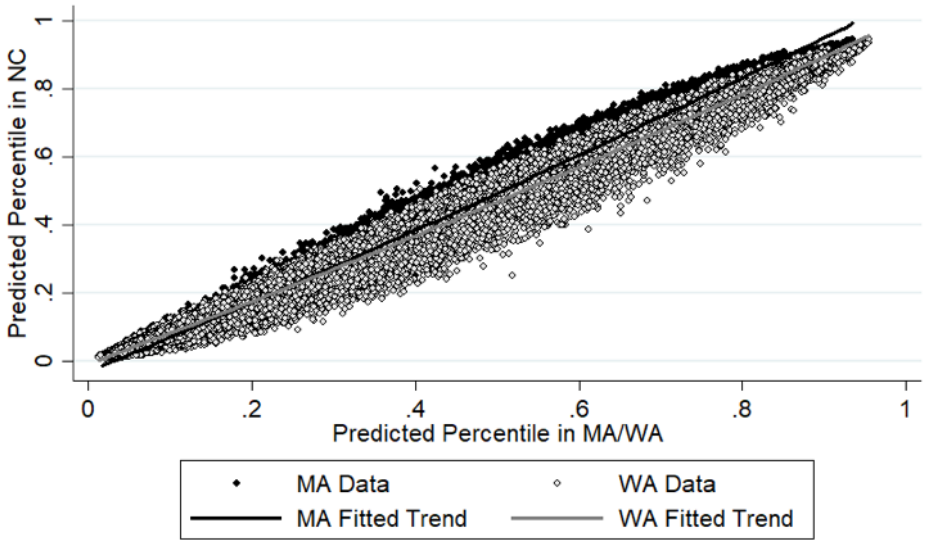
Notes: Scatterplot of predicted probability of scoring in the top half of 8th grade math tests in Washington compared to predicted probabilities in North Carolina and Massachusetts, estimated on students in Washington.

Figure A31: Scatterplot of Predicted Probabilities of Top Half 8th Grade Math Tests in MA vs Predicted Probabilities from Out-of-State Models (3rd Grade)



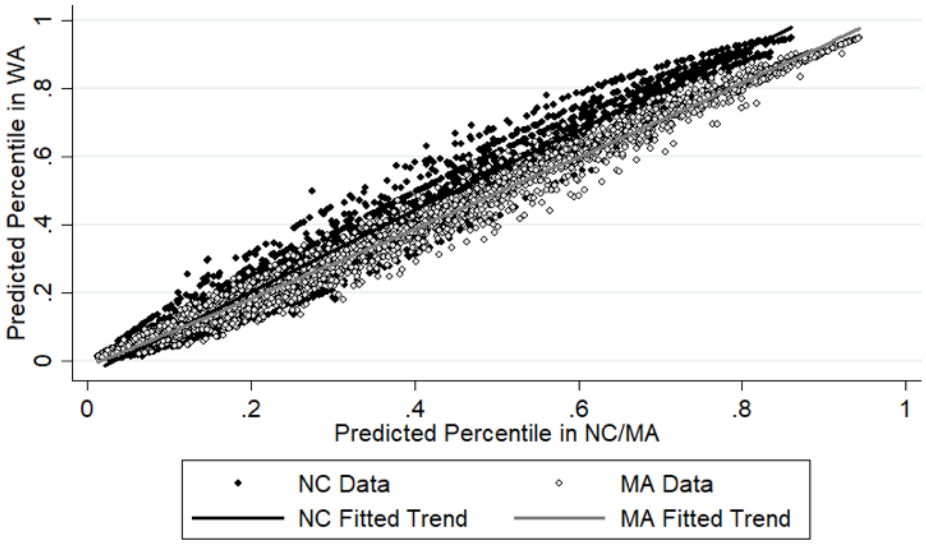
Notes: Scatterplot of predicted probability of scoring in the top half of 8th grade math tests in Massachusetts compared to predicted probabilities in North Carolina and Washington, estimated on students in Massachusetts.

Figure A32: Scatterplot of Predicted Probabilities of Top Half 8th Grade Math Tests in NC vs Predicted Probabilities from Out-of-State Models (3rd Grade)



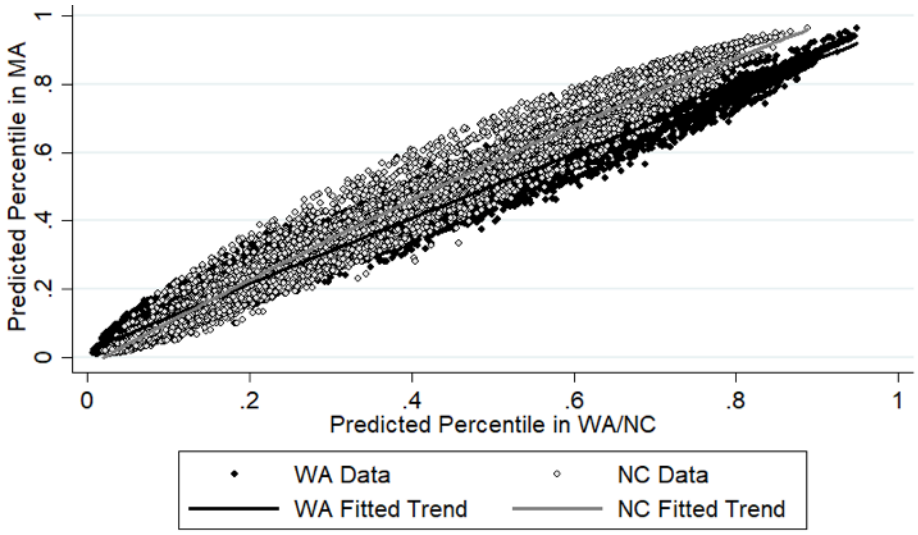
Notes Scatterplot of predicted probability of scoring in the top half of 8th grade math tests in North Carolina compared to predicted probabilities in Massachusetts and Washington, estimated on students in North Carolina.

Figure A33: Scatterplot of Predicted Probabilities of Top Half High School Math Tests in WA vs Predicted Probabilities from Out-of-State Models (3rd Grade)



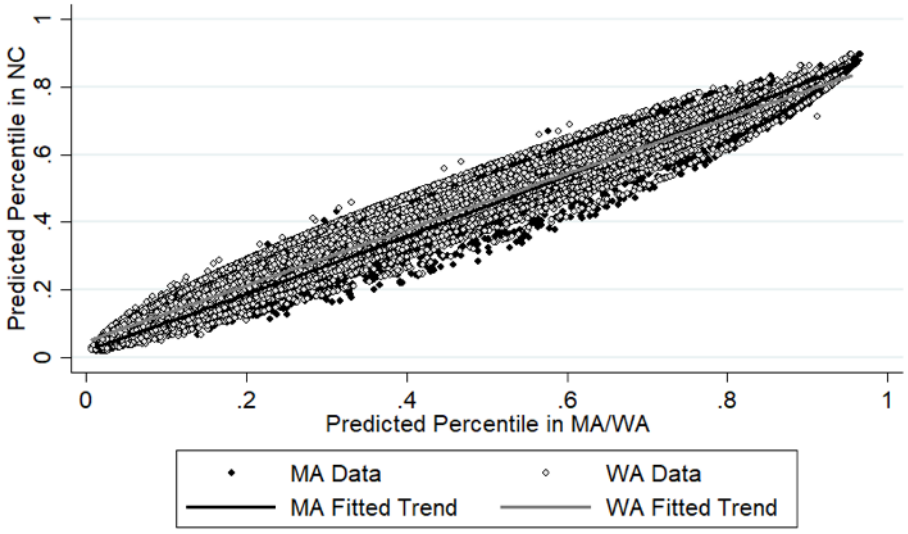
Notes: Scatterplot of predicted probability of scoring in the top half of high school math tests in Washington compared to predicted probabilities in North Carolina and Massachusetts, estimated on students in Washington.

Figure A34: 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 A35: 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 compared to predicted probabilities in Massachusetts and Washington, estimated on students in North Carolina.

Table A1: Model Coefficients of 8th Grade Math Percentile by Sample Attrition Status

	Overall		Massachusetts		North Carolina		Washington	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Math Percentile	0.539***	0.536***	0.489***	0.488***	0.56***	0.557***	0.505***	0.504***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
Reading Percentile	0.176***	0.173***	0.201***	0.200***	0.167***	0.163***	0.182***	0.181***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)
FRL	-5.315***	-4.903***	-6.94***	-6.296***	-4.891***	-4.493***	-5.847***	-5.556***
	(0.040)	(0.043)	(0.117)	(0.123)	(0.047)	(0.051)	(0.097)	(0.101)
Female	0.661***	0.496***	-0.415***	-0.456***	1.173***	0.949***	-0.21***	-0.286***
	(0.034)	(0.036)	(0.089)	(0.094)	(0.041)	(0.043)	(0.088)	(0.092)
White	(Reference Category)		-0.973***	-1.199***	-3.185***	-3.221***	-5.44***	-5.454***
			(0.873)	(0.932)	(0.208)	(0.222)	(0.267)	(0.285)
Amer. Indian	-3.640***	-3.587***	10.793***	10.683***	11.223***	10.791***	7.616***	7.472***
	(0.194)	(0.206)	(0.209)	(0.217)	(0.153)	(0.163)	(0.152)	(0.156)
Asian	10.239***	10.226***	-0.72***	-1.052***	-3.58***	-4.076***	-3.706***	-4.052***
	(0.296)	(0.314)	(0.183)	(0.193)	(0.053)	(0.056)	(0.192)	(0.202)
Black	-3.313***	-3.242***	-0.535*	-0.633*	-0.754***	-0.888***	-0.206	-0.129
	(0.285)	(0.304)	(0.309)	(0.325)	(0.131)	(0.14)	(0.307)	(0.322)
Multiracial	-0.236	0.117	-1.335***	-1.574***	2.393***	2.258***	-1.399***	-1.51***
	(0.267)	(0.284)	(0.156)	(0.165)	(0.105)	(0.113)	(0.14)	(0.146)
Hispanic	0.772***	0.868***	-0.888***	-0.852***	2.1***	2.045***	-3.086***	-3.305***
	(0.290)	(0.307)	(0.22)	(0.232)	(0.176)	(0.193)	(0.248)	(0.26)
Learning Disability	0.102	-0.045	4.258***	4.289	0.433***	0.071	3.402***	3.202***
	(0.114)	(0.122)	(0.193)	(0.203)	(0.142)	(0.152)	(0.17)	(0.177)
LEP	2.806***	2.593***	-7.109***	-7.159***	-6.266***	-6.134***	-3.958***	-3.978***
	(0.003)	(0.097)	(0.151)	(0.159)	(0.148)	(0.162)	(0.167)	(0.174)
Special Education	-5.170***	-5.319***						
	(0.087)	(0.092)						
No HS Sample Attrition		X		X		X		X
R	0.559	0.557	0.548	0.544	0.576	0.576	0.511	0.509
N	1,261,274	1,118,215	184,110	166,023	849,906	742,063	227,258	210,129

Notes: Regression model coefficients predicting 8th grade math tests, with and without students who exit in grades 9-12.

Table A2: Correlations of Predicted Advanced Course-Taking 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	0.798	1										
	3rd & 8th	0.860	0.991	1									
MA	3rd	0.939	0.746	0.808	1								
	8th	0.715	0.928	0.920	0.777	1							
	3rd & 8th	0.763	0.925	0.929	0.828	0.994	1						
NC	3rd	0.954	0.761	0.820	0.916	0.698	0.743	1					
	8th	0.765	0.960	0.951	0.737	0.911	0.906	0.807	1				
	3rd & 8th	0.829	0.953	0.961	0.800	0.902	0.911	0.869	0.991	1			
WA	3rd	0.903	0.709	0.774	0.943	0.722	0.772	0.884	0.708	0.774	1		
	8th	0.702	0.916	0.908	0.744	0.955	0.950	0.690	0.909	0.900	0.783	1	
	3rd & 8th	0.774	0.907	0.918	0.818	0.945	0.955	0.758	0.899	0.909	0.858	0.988	1

Notes: Correlations across state and grade level of test score used to predict advanced course-taking, for all states and each state individually.

Table A3: Correlations of Predicted Graduation 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	0.840	1										
	3rd & 8th	0.874	0.993	1									
MA	3rd	0.925	0.788	0.810	1								
	8th	0.798	0.929	0.920	0.868	1							
	3rd & 8th	0.791	0.928	0.917	0.863	0.999	1						
NC	3rd	0.946	0.793	0.825	0.882	0.756	0.749	1					
	8th	0.803	0.955	0.950	0.757	0.886	0.884	0.847	1				
	3rd & 8th	0.854	0.947	0.953	0.796	0.876	0.872	0.900	0.991	1			
WA	3rd	0.961	0.817	0.845	0.947	0.835	0.829	0.913	0.776	0.823	1		
	8th	0.808	0.971	0.964	0.798	0.959	0.959	0.768	0.926	0.918	0.843	1	
	3rd & 8th	0.798	0.970	0.962	0.789	0.958	0.956	0.757	0.925	0.915	0.835	0.999	1

Notes: Correlations across state and grade level of test score used to predict graduation, for all states and each state individually.

Table A4: 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: 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, free or reduced-price lunch status, and enrollment status in special education. The regressions labeled "Overall" control for state fixed effects.

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

Table A5: Model Coefficients of 8th Grade Math Percentile by State and Imputation Value

Panel 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, free or reduced-price lunch status, and enrollment status in special education.

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

Table A6: Model Coefficients by State with Imputation for High School Math Percentile (3rd Grade)

Panel A: Massachusetts

	UnImputed	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

	UnImputed	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

	UnImputed	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, free or reduced-price lunch status, and enrollment status in special education.

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

Table A7: Model Coefficients by State with Imputation for Advanced Course-Taking (3rd Grade)

Panel A: Massachusetts

	UnImputed (A1)	Imputed (A2)	+10% (A3)	+25% (A4)	-10% (A5)	-25% (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

	UnImputed (B1)	Imputed (B2)	+10% (B3)	+25% (B4)	-10% (B5)	-25% (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

	UnImputed (C1)	Imputed (C2)	+10% (C3)	+25% (C4)	-10% (C5)	-25% (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, free or reduced-price lunch status, and enrollment status in special education.

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

Table A8: Model Coefficients by State with Imputation for Graduation (3rd Grade)

Panel A: Massachusetts

	UnImputed (A1)	Imputed (A2)	+10% (A3)	+25% (A4)	-10% (A5)	-25% (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

	UnImputed (B1)	Imputed (B2)	+10% (B3)	+25% (B4)	-10% (B5)	-25% (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

	UnImputed (C1)	Imputed (C2)	+10% (C3)	+25% (C4)	-10% (C5)	-25% (C6)
3rd Grade Math Percentile	0.139***	0.114***	0.001***	0.001***	0.001***	0.001***
	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
3rd Grade Reading Percentile	0.115***	0.095***	0.114***	0.114***	0.113***	0.111***
	(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, free or reduced-price lunch status, and enrollment status in special education.

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