Adapting Course Placement Processes in Response to COVID-19 Disruptions

Guidance for Schools and Districts

JONATHAN SCHWEIG, ANDREW MCEACHIN, MEGAN KUHFELD,
LOUIS T. MARIANO, MELISSA DILIBERTI

Sponsored by the Institute of Education Sciences

For more information on this publication, visit www.rand.org/t/RRA1037-1

Published by the RAND Corporation, Santa Monica, Calif.

Preface

The novel coronavirus disease 2019 (COVID-19) pandemic has created an unprecedented set of obstacles for schools and exacerbated existing structural inequalities in public education. In spring 2020, as schools went to remote learning formats or closed completely, end-of-year assessment programs ground to a halt. As a result, schools began the 2020–2021 school year without student assessment data, which typically play a critical role in many districts' organizational processes. One such process concerns the selection of students for specialized programming or the placement of students into courses. Although many schools and districts rely on multiple sources of data to guide students' course placements, valid assessment-based information is seen as particularly valuable because it is predictive of future learning outcomes and can safeguard against potential bias that arises from more-subjective sources, such as teacher recommendations, ensuring that all students (and particularly underrepresented students) have equitable access to advanced courses.

Without such data, many school and district leaders developed ad hoc strategies to make course placements possible at the beginning of the 2020–2021 school year. In this way, the pandemic has created a great deal of ambiguity about how to accurately place students into courses. Important questions remain concerning whether these ad hoc strategies provide consistent and trustworthy information and whether these strategies preserve the goal of ensuring equity and mitigating disparities among black and Hispanic students, economically disadvantaged students, students with disabilities, English learners, and their white and more advantaged peers.

This report had two overall objectives. The first objective was to interrogate the consequences—intended and unintended—that may have arisen as a result of the strategies that schools and districts adopted to determine students' course placements for the 2020–2021 school year. We address this goal by comparing and contrasting three potential strategies and subsequently examining the ways in which the pandemic may have influenced the consistency of decisionmaking under these strategies, as well as the extent to which these strategies work equally well for all students, regardless of student or school demographics. The second objective was to articulate an investigatory framework that can be applied to a wide variety of assessment scenarios and that can guide and inform local decisionmaking by schools and districts. It is our hope that this report encourages schools and districts to systematically interrogate the equity implications of strategy adoption and to explore the potential of unintended negative consequences of particular strategies for specific kinds of decisions.

This report is the first of three that will examine the impacts of COVID-19-related assessment disruptions on school and district processes. Future reports will address strategic decisionmaking for accountability and evaluation purposes. Those reports will be released in summer 2021 and spring 2022.

This study was undertaken by RAND Education and Labor, a division of the RAND Corporation that conducts research on early childhood through postsecondary education programs, workforce development, and programs and policies affecting workers, entrepreneurship, and financial literacy and decisionmaking. The research reported here was sponsored by the Institute of Education Sciences, U.S. Department of Education, through R305U200006 to RAND. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

More information about RAND can be found at www.rand.org. Questions about this report should be directed to jschweig@rand.org, and questions about RAND Education and Labor should be directed to educationandlabor@rand.org.

Contents

APPENDIXES

Figures

Tables

Acknowledgments

The authors are grateful to the expert panelists who provided insight and guidance on this report: June Ahn, Sharon Bi, Betheny Gross, Laura Hamilton, Matthew Raimondi, Kevin Schaaf, Jessaca Spybrook, and Katharine Strunk. This document benefited substantively from feedback from Pete Goldschmidt and Dan Goldhaber. Emily Ward provided expert editing. Any flaws that remain are solely the authors' responsibility.

Abbreviations

COVID-19	coronavirus disease 2019
ESSA	Every Student Succeeds Act
FRPL	free or reduced-price lunch
MSE	mean square error
OLS	ordinary least squares

CHAPTER ONE

Introduction

The novel coronavirus disease 2019 (COVID-19) pandemic has created an unprecedented set of obstacles for schools and exacerbated existing structural inequalities in public education. In spring 2020, the pandemic shut down school buildings across the world. According to reports from the United Nations Educational, Scientific and Cultural Organization, such school closures ultimately affected over 90 percent of the world's student population (Giannini, Jenkins, and Saavedra, 2020). In the United States, all 50 states closed schools to in-person instruction at some point in spring 2020 (Peele and Riser-Kositsky, 2020).

Forecasts based on historical learning trends and the literature on summer learning loss suggest the possibility that the pandemic will have drastic impacts on student learning (e.g., Kuhfeld, Soland, et al., 2020). Although it will take years to unpack how the pandemic affected students' learning and social and emotional development, early data from fall 2020 suggest that mathematics achievement was about 5–10 percentile points lower this year, as compared with same-grade students in the prior year (Kuhfeld, Tarasawa, et al., 2020), and these data may understate the extent of the problem given that many more-vulnerable students did not participate in fall testing (Johnson and Kuhfeld, 2020). Student course grades for the first period of the 2020–2021 academic year suggest a surge in failure rates across the country (Strauss, 2020). In addition to research on the overall impact of the pandemic on student learning, careful research will be needed to understand the extent to which the pandemic has exacerbated long-standing disparities in opportunity and achievement between students of color, economically disadvantaged students, English-language learners, and students with disabilities and their white and more advantaged peers.

In spring 2020, as schools went to remote learning formats or closed completely, end-of-year assessment programs—including annual statewide accountability tests (such as those required by the Every Student Succeeds Act), interim assessments, and benchmark assessments (such as the NWEA MAP Growth assessment and ACT Aspire)—ground to a halt. As a result, schools began the 2020–2021 school year without any end-of-year student assessment data, which may play a critical role in local decisionmaking. Without such data, many school and district leaders were forced to develop ad hoc strategies to make such organizational processes possible at the beginning of the 2020–2021 school year. State and national professional organizations developed principles to encourage schools and districts to make the most of the data they had available and to ensure, to the best extent possible, that students were held harmless for the impacts of the pandemic (Marion et al., 2020). However, schools and districts often did not have time or resources for a full exploration of the intended and unintended consequences of these strategies.

End-of-year assessment data are incorporated into school and district policy- and decisionmaking in a variety of ways. Such data are used to evaluate school or district improvement efforts (Beaver and Weinbaum, 2013; Hamilton et al., 2009), to support instructional planning, and to monitor achievement gaps between student subgroups. Schools and districts also rely on end-of-year assessment data, in conjunction with course grades and teacher recommendations, to make decisions regarding students' course placements and select students for specialized programming, including advanced mathematics or English language arts classes (Beaver and Weinbaum, 2013; Means, Padilla, and Gallagher, 2010). Valid test-based information is widely seen as an important component of course placement recommendations because it is predictive of future learning outcomes (Goldhaber, Wolff, and Daly, 2020; Huang, Snipes, and Finkelstein, 2014; Kriegler and Lee, 2006) and because it can safeguard against potential bias that arises from more-subjective sources, including teacher recommendations, ensuring that all students (and particularly underrepresented students) have equitable access to advanced courses. Course placement processes that rely exclusively on subjective sources of information, such as teacher recommendations, contribute to persistent racial disparities in representation in advanced academic programs (Card and Giuliano, 2016; Grissom and Redding, 2016; Oakes, 1985; U.S. Department of Education Office for Civil Rights, 2014). In fact, to safeguard against the systematic bias of such subjective measures as teacher recommendations, several states have legal mandates that require that assessment information be included in course placement processes (Gao and Adan, 2016; RCW 28A.320.195).

This report had two overall objectives. The first objective was to interrogate the consequences—intended and unintended—that may have arisen as a result of the strategies that schools and districts adopted to determine students' course placements for the 2020–2021 school year. We examine whether these strategies preserve the goal of ensuring equity and mitigating disparities among black and Hispanic students, economically disadvantaged students, and their white and more advantaged peers. We specifically investigate two research questions that are related to this objective:

1. To what extent can older assessment information be used to ameliorate the problem of missing test data used for course placements?
2. To what extent do student and school characteristics influence the consistency of using older data in course placement strategies?

We investigate these research questions by contrasting three widely used or recommended strategies for addressing missing data: (1) a simple replacement strategy, which uses students' scores on their most recently available assessment in place of their end-of-year assessment scores; (2) a regression-based replacement strategy, which uses a regression model based on past student achievement trends to predict future end-of-year assessment scores; and (3) a multiple replacement strategy, which employs a multiple imputation framework. We examine the consistency of these methods by comparing the course placement recommendations using actual test scores with the course placement recommendations using replacement test scores. We then use school and student characteristics to explore the extent to which there might be differences in consistency that are associated with student race, gender, and school poverty. We find that the consistency of these strategies varies substantially from district to district. In particular, we find that substantial misclassification may occur in districts in which subgroups of students have significant changes in achievement or content mastery between testing periods. We also

find that using regression-based replacement strategies that do not account for differences in students' school experiences may compromise the consistency of course placement recommendations. In particular, because these models do not recognize that black, Hispanic, and economically disadvantaged communities systematically have access to lower-quality schools than their white, Asian, and more advantaged peers, they may overestimate test scores for black and Hispanic students or students who attend high-poverty schools and underestimate test scores for white students, Asian students, and students who attend more-economically advantaged schools.

Course placement is just one of many decisionmaking processes that were interrupted or altered by the pandemic. Thus, our second objective was to articulate an investigatory framework that can be applied to a wide variety of assessment scenarios and that can guide and inform local decisionmaking by schools and districts. It is our hope that this report encourages schools and districts to systematically interrogate the equity implications of strategy adoption and to explore the potential of unintended negative consequences of particular strategies for specific kinds of decisions.

This remainder of this report is organized into four chapters and three appendixes. In Chapter Two, we describe the role of end-of-year assessments in course placement processes and provide examples of district policies. In Chapter Three, we detail our research questions and describe our sample and methods. In Chapter Four, we present the results of our analyses and offer recommendations for schools and districts. In Chapter Five, we conclude with a discussion of the particular implications of our findings for decisionmaking during the COVID-19 crisis. Appendix A provides descriptive information and details on inclusion criteria for our analytic sample, Appendix B provides technical details on the analytic methods we used, and Appendix C provides complete results from our analyses.

CHAPTER TWO

Study Background and Context

Common end-of-year assessments have been administered as a part of U.S. public schooling for nearly a century (McDonnell, 1994). The current assessment context was spurred by the 2001 No Child Left Behind (NCLB) mandate, which required states to implement annual assessment systems with aggregated reporting for the whole school and for different subgroups of students, such as English-language learners, racial minorities, students in poverty, and students enrolled in special education (Haertel and Herman, 2005; Klein, 2015). The implementation of the Every Student Succeeds Act (ESSA) in 2015 brought about some substantive changes to NCLB-era policies; under the new law, states needed to incorporate both academic and nonacademic indicators into their assessment and accountability frameworks. In most (if not all) state plans, these academic indicators include proficiency on state tests, English-language proficiency, and student growth measures that can be broken out by subgroup (Klein, 2016; Olson, 2019).

Although policy research historically has paid a great deal of attention to the accountability and evaluation purposes of assessment, end-of-year assessments serve many purposes and support decisionmaking by a variety of stakeholders. Parents and guardians might use end-of-year assessments to identify a child's academic strengths and weaknesses, to monitor whether a child met state standards (Eissenberg and Rudner, 1988), or to evaluate school quality (Favero and Meier, 2013). School leaders might use assessment data to evaluate whether students are ready for grade advancement or graduation, to evaluate school quality, or for instructional planning (California Department of Education, 2020; Dougherty, 2015; Hamilton et al., 2009; Woods, 2017). District and state leaders might use assessment data to evaluate interventions, to inform professional development, or to allocate school resources. Schools and districts also rely on end-of-year assessment data to inform students' course placement recommendations and to select students for specialized programming (Beaver and Weinbaum, 2013; Louis et al., 2010; Means, Padilla, and Gallagher, 2010), which is the focus of this report.

An Overview of Student Placement Decisions in Policy and Practice

According to a nationally representative survey (Means, Padilla, and Gallagher, 2010), the vast majority of public-school districts in the United States (84 percent) rely on end-of-year student assessment scores—often in conjunction with course grades and teacher recommendations—to make determinations for students' course placements. There is also research suggesting that end-of-year assessment data are used by schools to group students by performance level (i.e., for "tracking"; see Beaver and Weinbaum, 2013) and for other placement decisions, including

screening for gifted and talented programs, placement decisions for students moving into the school district, screening for students to enroll in accelerated courses, enrollment in algebra I, and enrollment in Advanced Placement courses.

The majority of student placement policies are focused on middle and high school–aged students, and a particular amount of attention is paid to enrollment in advanced mathematics courses (including algebra I) (e.g., Cortes, Goodman, and Nomi, 2015; Gamoran, 1992; Kelly, 2007; McEachin, Domina, and Penner, 2020; Rickles, 2011; Stein et al., 2011). However, other kinds of placement decisions, including decisions around eligibility for gifted and talented programs, often are also relevant to elementary grades (Bui, Craig, and Imberman, 2014; Grissom and Redding, 2016; McClain and Pfeiffer, 2012).

Most course placement processes are governed by a belief that decisions of how to allocate students to selective academic programs should be informed by evidence that students will be successful in such programs. Current academic performance and content mastery are widely regarded as the most relevant predictors of future performance (Archbald, Glutting, and Qian, 2009). End-of-year assessment scores have been shown to be effective indicators of likely success in future classes (Anderson and Newell, 2008; Goldhaber, Wolff, and Daly, 2020; Huang, Snipes, and Finkelstein, 2014; Kriegler and Lee, 2006)

The increased reliance on assessment scores as a basis for course placement decisions is based on historical evidence that course placement processes that rely on subjective information, including teacher recommendations, can be systematically biased against certain groups of students. Prior research documents that, on average, white teachers have lower expectations for students of color than for white students (e.g., Gershenson, Holt, and Papageorge, 2016; Papageorge, Gershenson, and Min Kang, 2020). These biases manifest in the underplacement of black, Hispanic, and economically disadvantaged students into advanced academic programs relative to their white peers (e.g., Figlio, 2005; Gamoran, 1987; Grissom and Redding, 2016; Oakes, 1985). These disparities in access to advanced courses exacerbate other educational disparities, including college readiness and college admission (Gao, 2016; Jaschik and Lederman, 2018; National Academies of Sciences, Engineering, and Medicine, 2019). Means, Padilla, and Gallagher, for example, highlight a case study of a school district that used teacher recommendations as the primary criteria for middle school algebra placement and found that black students were systematically underenrolled in algebra courses. The district changed its course recommendation process to include a math assessment and found that "200 students we assigned to Algebra [using this approach] would not have been recommended by teachers" (Means, Padilla, and Gallagher, 2010, p. 25).

Unlike teacher recommendations, assessment scores are assumed to be objective measures of a student's understanding, a student's ability, or whether a student has met or exceeded grade-level standards for learning. Several states have policies that explicitly require that assessments be used to inform course placement and enrollment in gifted and talented programs. For example, California recently codified the use of transparent and fair placement policies for students in grades 9–12 in Senate Bill 359 (California State Senate, 2015). The law aimed to improve equity in the math placement process, and, after its passage, over 90 percent of districts in California indicated that they use test scores as a component in determining appropriate math courses for students (Gao and Adan, 2016). In Washington state, common school provisions state that assessment results must be used as the sole criteria for academic acceleration for high school students. In addition, any student who meets or exceeds the state standard on the eighth grade or high school English language arts or mathematics statewide student

assessment must be automatically enrolled in the next-most-rigorous level of advanced courses or program offered by the high school (RCW 28A.320.195).

Many of the systems that incorporate students' assessment scores into course placement decisions do so using a decision rule that employs a cut score or a minimum threshold score of some kind. Sometimes, these scores are norm-referenced—i.e., a student's performance is compared with those of other students in a reference population; other times, these scores are criterion-referenced—i.e., a student's performance is compared with grade-level standards or learning goals (National Research Council, 2002). As one example of a criterion-based course placement rule, Broward County Public Schools, the sixth-largest public-school district in the United States, requires students to have attained a performance level of 3 or higher (indicating satisfactory performance) on the eighth-grade Florida Standards Assessment to be placed into honors English in high school (School Board of Broward County, Florida, 2019). Similar systems are used for eighth-grade algebra placements in California school districts, including the Los Angeles Unified School District, where proficiency levels on sixth-grade math tests are used as the primary source of information for placement into accelerated mathematics pathways and algebra I (Huang, Snipes, and Finkelstein, 2014; Kelly, 2007; Los Angeles Unified School District, 2017; McEachin, Domina, and Penner, 2020).

Other school districts rely on norm-referenced tests for course placement decisions. For example, Chicago Public Schools uses NWEA's MAP assessment to make course placement determinations. Specifically, the district requires that students score in the 95th percentile on the NWEA MAP for two consecutive testing seasons in the subject area in which acceleration is sought, in addition to a minimum grade point average of 4.0 and other assessment information (Chicago Public Schools, undated). Furthermore, McClain and Pfeiffer, 2012, found that 18 states have policies that stipulate specific test scores for students to qualify as gifted and that Mississippi has a specific requirement that students score at or above the 90th percentile on an achievement test to be placed into gifted programming. Norm-referenced scores can also be used to support course placement processes in which only a certain portion of students can be placed into advanced classes; for example, Card and Giuliano, 2016, describes a school district with a mandate to establish separate classrooms for gifted fourth- or fifth-graders. Because only a small number of students are identified as gifted, the remaining seats in these classes are populated with the highest-achieving students based on prior test scores.

How Have Schools and Districts Responded to the Challenge of Missing Assessment Data for Course Placement?

In spring 2020, in response to the COVID-19 pandemic and widespread school closures, the U.S. Department of Education granted waivers to all 50 states to allow state and local education agencies to bypass annual ESSA testing requirements (Gewertz, 2020). Additionally, most school districts decided to skip interim or benchmark assessments that normally would be administered in the spring (for MAP Growth specifically, less than 5 percent of students who are normally assessed in the spring were administered tests in spring 2020). As a result, the pandemic has created a great deal of ambiguity about how to select students for specialized programming and accurately place students into courses.

In some ways, schools and districts are simply reckoning with an enormous and complex missing data problem. Districts have had to deal with missing assessment data issues before.

For example, between 2010 and 2014, interruptions occurred during the online administration of statewide assessments in Florida, Indiana, Kentucky, Minnesota, Oklahoma, and Wyoming (Martineau et al., 2015). In 2015, similar testing interruptions occurred in Georgia, Montana, Nevada, and North Dakota (Martineau et al., 2015). In some states, interruptions were pervasive enough that states needed to invalidate assessment scores altogether. For example, in 2016, Tennessee experienced such widespread issues with the transition to an online standardized test that the Tennessee Department of Education suspended annual statewide assessments for all students in grades 3–8 (Data Quality Campaign, Alliance for Excellent Education, Collaborative for Student Success, 2020). However, there are many ways in which the current missing data problem is unlike any other missing data problem previously encountered, in terms of both its scale and the way in which it is inextricably linked to a disruption in formal schooling.

That said, we identified three primary strategies that schools and districts might be using to deal with missing student assessment data and to move forward with course placement decisions for the 2020–2021 school year. Broadly speaking, all three strategies can be classified as score-replacement methods, in which an available student score is used as a substitute or a surrogate for an unavailable score. In statistics, these substitution or replacement processes are referred to as various forms of imputation. The three broad classes of score-replacement strategies used in this report are summarized in Table 2.1 and organized in order of increasing complexity and technical sophistication. While Appendix B provides more-technical details of these imputation methods, Table 2.1 provides a brief description of each method to highlight important features and summarize literature describing how schools and districts have used such strategies in the past. We also summarize some of the key underlying assumptions of each method.

Simple Replacement

In the absence of end-of-year assessment scores in spring 2020, school districts may have opted to use available data from previous assessments to make initial recommendations for students' course placements for the 2020–2021 school year. Such a replacement strategy can take several

Table 2.1
Three Categories of Replacement Methods

Method	Description	Key Underlying Assumptions
Simple replacement	Use students' scores on most recent available assessment in place of their end-of-year assessment scores	Assumes that students' percentile rankings or content mastery are unlikely to change dramatically between assessment administrations (e.g., students who were at the 65th percentile in winter 2020 are likely to be at or very near the 65th percentile in spring 2020)
Regression-based replacement	Use a regression model based on past student achievement trends (and demographic information) to predict missing end-of-year assessment scores	Assumes that relationships among students' scores in 2019–2020 are the same as in prior years. Usually assumes that students encounter an "average schooling experience" in the future
Multiple replacement	Use a multiple imputation framework to generate several plausible replacement scores for each student that allows for some uncertainty about students' missing assessment scores	Assumes that relationships among students' scores in 2019–2020 are the same as in prior years

forms. For school districts that administer only the end-of-year (spring) assessments required by ESSA, using older assessments to inform placement decisions might mean using student assessment data from all the way back at the end of the previous school year (spring 2019). For school districts that use standardized interim or benchmark assessments that are administered at multiple points throughout the school year (for example, the NWEA MAP Growth assessment, Renaissance Math, ACT Aspire, and Mathematics Assessment Resource Service assessments), using older assessments to inform placement decisions might mean using assessment data from the fall or winter of the 2019–2020 school year. It is also possible to use scores on a benchmark or interim assessment as a substitute for an end-of-year assessment. Although we are unable to describe the prevalence of this strategy, recent work on 2020 achievement tests noted that several state officials were intending to use such a strategy (Center for Research on Education Outcomes, 2020), and our own conversations with assessment directors and research coordinators at local education agencies corroborate this finding. As one concrete example, the Santa Monica Malibu–Unified School District (SMMUSD) based recommendations for seventh-grade mathematics course placements in 2020–2021 on fifth-grade California state assessment scores. Specifically, students needed to exceed fifth-grade standards to be recommended for accelerated courses (SMMUSD, 2020).[1]

There are several assumptions underlying the simple replacement strategy that depend on whether course placement is a normative process or a criterion-referenced process and on the assessment content. For normative score uses, a key underlying assumption of this strategy is that student percentile ranking does not decrease substantially between test administrations. That is, students who are at or above a given percentile in 2019, for example, would be unlikely to fall below the cutoff in a meaningful way in 2020 (Center for Research on Education Outcomes, 2020). Correspondingly, for criterion-referenced decisions, the assumption would be that students who exceed standards in 2019 would be likely to exceed standards for learning in 2020.

Simple replacement strategies also assume that scores on one test mean the same thing about student achievement mastery as scores on another test (and can be interpreted equivalently). Although this assumption might be suitable if the substituted assessment is similar in format and score (for example, substituting one end-of-year assessment for another end-of-year assessment), it might be more complicated in other cases, such as if schools and districts substitute benchmark or interim assessments for end-of-year state assessments. Assessments are designed to support specific decisions about test takers and to support policy analysis, program evaluation, and educational accountability (Kane, 2013). Not all assessments are designed to support all of these purposes (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014; Gong, 2020; Popham, 2016). End-of-year assessments might be designed to support claims about content mastery in a particular subject at a particular grade level; the content of these assessments might be selected to broadly survey the year's topics, curriculum standards, and competencies. Interim assessments, on the other hand, might be designed with a diagnostic focus for specific skills and content and might support claims about the extent to which students are prepared to advance to the next instructional unit. A significant amount of evidence would need to be gathered to support the assumption that test scores from different assessments can be inter-

[1] SMMUSD typically uses multiple measures for course placement, including sixth-grade course grades.

preted equivalently (Kane, 2006; Messick, 1995; Shepard, 1993) and thus to justify the use of such a replacement strategy.

One other consideration for the simple replacement strategy is that it assumes that there are contiguous assessment results that are available for students. This is a reasonable assumption for elementary and middle school grades but might be less practical in high school contexts, in which testing information might not be available in every grade.

Regression-Based Replacement

Replacement scores can also be obtained using a regression model in which prior achievement data are used to estimate a relationship that characterizes students' year-to-year growth. The regression model either can use a straightforward ordinary least squares (OLS)–based specification or can incorporate school-to-school variability in average student achievement by incorporating what are known as school *random effects*. In either case, the method is generally implemented in two steps. In the first step, the regressions are applied to historical data to obtain model parameters that quantify the associations among the predictors and outcomes. In the second step, these estimated model parameters are used to make out-of-sample predictions for future test scores. Although all of the regression models can incorporate a variety of variables into score prediction, including student and school characteristics, there are important trade-offs between equity and consistency that we discuss in further detail in Chapter Four. Equations for these models are included in Appendix B.

Several states, including Ohio, Pennsylvania, and Tennessee, use models of this kind to make student score predictions (Pennsylvania Department of Education, 2020; SAS, 2017), and, in some districts, these score projections are used as a component of course placement processes. For example, in Knoxville public schools, course placement recommendations are based in part on projected state testing percentiles (LaRoy and Roberts, 2019). Other districts in Tennessee use projected percentiles on algebra I end-of-course assessments to inform seventh-grade algebra placements (Williamson County Schools, undated).

Regression-based replacement strategies assume that students' future test scores are a function of their prior test performance and possibly student and school demographics. Under the assumption that the relationship between students' current test scores and their prior test scores and demographics holds over time, we can use this relationship to project how a student might score in a future year (e.g., spring 2020). Such models typically assume that future academic achievement follows the same linear trajectory as past academic achievement. However, it is possible to adjust these models to accommodate nonlinear achievement growth. Specific models might differ in how they handle the nested nature of the data (e.g., students are embedded in classes, schools, and districts), the use of multiple years of prior achievement, and often the level of the score projection (e.g., these models can make projections not only for students but also for higher levels, such as teachers or schools).

All of the regression-based replacement methods also make the assumption that a student's current test score can be modeled as a function of her prior test scores and, in some cases, student and school demographics. Additionally, a key assumption in these regression-based replacement models concerns how the models account for the effect of schools on student learning. Some models ignore the fact that students attend schools of varying quality, and other models require users to make an assumption about the role of schools to generate individual student predictions. Options generally include (1) assuming that student growth is independent of the school that a student attends, (2) assuming that students will encounter an

average school experience, and (3) assuming that a school's contribution to student learning is an important part of the student's projected score (e.g., students who attend schools with positive impacts on student achievement will have higher projected scores, all else equal, than students who attend schools with negative impacts on student achievement).

A final note on regression-based replacement methods is that decisions about whether to include demographic covariates (such as a race and ethnicity) in score prediction models ultimately can be influenced by many factors. Two factors are particularly important. First, state and federal policy (e.g., ESSA) might not allow the use of student demographics if growth models are used to generate student predictions. Second, it is often more difficult to communicate the results of growth models that use student demographics. On the one hand, it might be important to account for factors that are outside the control of teachers or schools. On the other hand, families, students, teachers, and other stakeholders might interpret the use of such characteristics as an explicit and/or implicit decision to set different standards for different students based on their race or ethnicity.

Multiple Replacement

The simple replacement strategy and the regression-based replacement strategy that we have just described are used to create a single replacement score for each student. In doing so, these methods do not account for any uncertainty about the suitability of these replacement scores. For example, in the simple replacement case, it is assumed that the available data are an exact substitute for the missing data. In the regression-based replacement case, it is assumed that all of the relationships are estimated perfectly in the first stage. However, this is almost certainly not the case, because there is most likely some uncertainty around how well these replacement scores serve as a proxy for the actual scores. The multiple replacement strategy attempts to account for this uncertainty by using an iterative process in which multiple replacement values are generated for each student. A variety of complex statistical methods can be used to generate these replacement values, but, in general, such methods are referred to in the statistical literature as *multiple imputation* methods (e.g., Peugh and Enders, 2004). In response to prior testing interruptions (such as those caused by the adoption of online assessments described earlier), some states recommended exploring the potential for multiple imputation to address missing data in statewide accountability systems (Marion and Domaleski, 2012).

An important assumption underlying both the regression-based replacement and multiple replacement methods is that the relationships among scores in 2019–2020 are the same as in prior years, enabling the extrapolation of prior relationships to impute the missing scores. We consider this assumption further in Chapter Five. Additionally, these methods assume that the missing observations are missing randomly, after observed student characteristics are accounted for.

Other Important Considerations

Accurate Course Placement Has Important Consequences for Student Learning

As described previously, there are other potential use cases that could have served as the basis for this report. However, research suggests that accurate course placement plays an important role in school success, college readiness, and college admissions (Gao, 2016; Jaschik and Lederman, 2018; National Academies of Sciences, Engineering, and Medicine, 2019). Misplacement, either above or below the appropriate level, can have significant negative consequences for student learning. For example, students who are placed in courses in which they are unable

to keep pace and students who have weaknesses in key foundational areas frequently face significant difficulties reaching proficiency later on (Finkelstein et al., 2012; Stein et al., 2011). Conversely, students who are placed in courses that are too easy for them miss out on more-advanced educational opportunities.

If there are differential or drastic effects of COVID-19-related interruptions on students' prerequisite skills, misplacement for the 2020–2021 school year might be more pronounced as a result. Although we are not able to make causal claims about the direct impacts of course placement on student success, recent reports suggest that course failure rates in some districts were significantly higher when students reentered school in fall 2020 than in previous school years, particularly among remote learners, providing some preliminary evidence that some students may have been placed inappropriately this school year (Sawchuk, 2020; Strauss, 2020).

There Is Uncertainty Surrounding How Long COVID-19 Will Continue to Disrupt Standard Educational Practices, Systems, and Routines

Because of uncertainty regarding how long the pandemic will persist, the extent to which districts are in need of a one-off or short-term strategy to deal with missing assessment data for spring 2020 specifically, or a more robust strategy to deal with missing (or perhaps lower-quality) student assessment data spanning multiple school years, is currently unclear. Although all 50 states and the District of Columbia closed their schools to in-person instruction at some point during the 2019–2020 school year, school districts adopted a variety of reentry scenarios for fall 2020, which influenced their instructional and assessment practices for the 2020–2021 school year. According to Ballotpedia, as of November 18, 2020, nine states had state-ordered regional school closures for certain grade levels or allowed hybrid instruction only. Four states had state-ordered in-person instruction, and 37 states had reopening plans that varied by school or district (Ballotpedia, undated). In response to local public health guidelines, many of these reopening plans are under constant review, and, in some districts, there might even be some fluctuation between fully remote and hybrid systems. For example, on November 19, New York City Public Schools returned to remote instruction because of surging local COVID-19 infection rates; however, on December 7, the district began returning some students to in-person instruction.

Several recent conceptual papers offering policy guidance have recommended that schools and districts use substitute data sources to guide decisionmaking in fall 2020, including using short-turnaround assessment data to make course placement decisions and create instructional plans (e.g., Bellwether Education Partners, 2020). But as school closures persist through the 2020–2021 school year, more guidance is needed about how to proceed with student assessment data collected this school year. New research has raised important questions about the extent to which data are complete, consistent with prior administrations, and reflective of student understanding (Boyer and Keng, 2020). An examination of remote testing in fall 2020 found that students who took diagnostic assessments at home in fall 2020 showed an improvement over previous years while students who tested at school showed losses (Huff, 2020). These results raise concerns about the potential that some parents or guardians might be assisting in the assessment process during remote testing.

CHAPTER THREE

Study Approach

This report addresses the following research questions:

1. To what extent can older assessment information be used to ameliorate the problem of missing test data used for course placements?
2. To what extent do student and school characteristics influence the consistency of using older data in course placement strategies?

This report uses data from NWEA's anonymized longitudinal Growth Research Database to examine how course placement decisionmaking may have been affected by the lack of end-of-year assessment data from spring 2020. The Growth Research Database contains MAP Growth reading and math test scores, which are standards-aligned and norm-referenced. MAP Growth assessments are linked to most state accountability assessments (NWEA, 2020), allowing educators to use MAP Growth results across the school year to predict a student's likelihood of being proficient on the end-of-year state accountability tests. Although the MAP Growth assessments differ in many ways from state accountability tests in content, administration, and score interpretation, many districts administering MAP assessments use the scores in course placement processes, making them suitable for our analyses (Ryan, 2019).

In this chapter, we first briefly describe the data that we used for this study. We then describe the details of our research design. We conclude with a brief overview of limitations to the study.

NWEA Growth Research Database

NWEA's Growth Research Database contains MAP Growth reading and math test scores for over 7 million kindergarten through eighth-grade students within approximately 15,000 traditional public schools and 1,300 charter schools in all 50 states and the District of Columbia. NWEA collects information about student race, ethnicity, and gender. School-level characteristics (including the proportion of students eligible for free or reduced-price lunch [FRPL]) were obtained from the 2017–2018 Common Core of Data file from the National Center for Education Statistics. Overall, the sample closely aligns to the characteristics of U.S. public schools, including percentage of FRPL receipt; percentage of urban, rural, and suburban schools; and percentage of white students. However, the NWEA sample has a slight underrepresentation of Hispanic students compared with the national population. A comparison of the schools in our sample relative to the U.S. population of public elementary and middle schools is provided in Appendix A.

The MAP Growth assessments are computer-adaptive; are vertically scaled across grades; and measure student achievement in math, reading, language usage, and science using items aligned to state standards. This scaling allows educators to make various comparisons across time and space. MAP Growth is used for various purposes within and across districts, including as a measure of academic growth and student goal-setting, for curricular and programmatic placement decisions, as a component of admissions decisions for selective-enrollment high schools, as a universal screener for intervention programs, and as a component of teacher evaluation and school accountability systems. MAP Growth assessments are administered at multiple time points during the school year. Students typically take the NWEA MAP Growth exams in the fall, winter, and spring of each school year.

This report uses assessment data from four academic years (2016–2017, 2017–2018, 2018–2019, and 2019–2020). To compare and contrast the three categories of replacement strategies described in Chapter Two, we limited our analytic sample to include students who had spring assessment scores in both reading and math for school years 2016–2017, 2017–2018, and 2018–2019. To ensure that comparisons across replacement strategies are not based on differences in the analytic samples, we limited our sample to students who had complete student-level demographic information (race and ethnicity, gender, age, and grade level) and were matched to a school with a valid National Center for Education Statistics identification number (complete inclusion criteria for our analytic sample are included in Appendix A). Our final analytic sample were approximately 50 percent male, 50 percent white, 15 percent black, and 19 percent Hispanic (Table 3.1). See Appendix A for more details on the NWEA MAP sample schools and students.

Research Design

In this section, we describe the framework that we used to investigate our two research questions.

Research Question 1: Investigations of Replacement Strategies in Course Placement Decisions

To investigate the first question, which concerns the factors that schools and districts should consider when determining whether a particular imputation (replacement) strategy is appropriate for course placement decisions, we conducted a series of analyses using NWEA MAP data from the 2016–2017, 2017–2018, and 2018–2019 academic years. We focus on this range of school years because spring 2019 is the most recent spring term in which most students were assessed on MAP Growth. The purpose of our analyses is to compare and contrast the outcomes of various replacement methods. The primary framework for this investigation follows prior work in estimating consistency and accuracy of classifications based on test scores (e.g., Douglas and Mislevy, 2010; Livingston and Lewis, 1995). As described in Chapter Two, course placement often is based (at least in part) on policies that establish a cut score or a minimum percentile rank. In some districts, this score might be as high as the 90th percentile; in other districts, this cut score might be closer to the median. Even within districts, these cut scores can vary; for example, placement into algebra might be based on a lower cut score than placement into a gifted and talented program.

Table 3.1
Descriptive Statistics for the Study Sample

Demographic Characteristic	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
				Math				
White	0.51	0.51	0.50	0.49	0.49	0.49	0.50	0.50
Black	0.16	0.16	0.15	0.15	0.15	0.15	0.15	0.15
Hispanic	0.16	0.16	0.19	0.19	0.20	0.20	0.20	0.19
Asian	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Male	0.51	0.51	0.51	0.51	0.51	0.50	0.51	0.51
N	256,531	300,096	379,707	406,187	345,089	309,638	282,976	2,280,224
				Reading				
White	0.51	0.51	0.50	0.49	0.49	0.49	0.50	0.50
Black	0.16	0.16	0.15	0.15	0.15	0.15	0.15	0.15
Hispanic	0.16	0.16	0.19	0.19	0.20	0.20	0.20	0.19
Asian	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Male	0.51	0.51	0.51	0.51	0.51	0.50	0.51	0.51
N	256,536	300,044	379,688	406,182	345,070	309,627	282,985	2,280,132

NOTE: *N* = number of students.

We used the following course placement scenario for our investigations: Students must have a minimum score at the 65th percentile on their end-of-year assessment relative to the national population of MAP test takers to be recommended for higher-level courses. For example, to be recommended for placement into eighth-grade algebra or honors English, students would need to have an end-of-year assessment score at the 65th percentile in that subject. Our empirical investigation of each strategy had four steps:

1. Record a student's observed spring 2019 MAP score.
2. Using our three replacement strategies, generate a substitute spring 2019 MAP score. These replacement strategies are
   a. simple replacement: use spring 2018 assessment
   b. regression-based replacement: regress spring 2018 on spring 2017, and use regression results to predict spring 2019
   c. multiple replacement: impute spring 2019 based on whichever assessment scores were observed in 2016–2017 and 2017–2018.
3. Apply our cut-point criterion to the actual spring 2019 MAP scores and the replacement spring 2019 MAP scores, and make a course recommendation for the sets of actual and replacement scores.
4. Examine the congruence between recommendations based on the actual and replacement spring 2019 scores to determine classification consistency.

All of these analyses were conducted separately for math and reading MAP scores and separately by grade level. The three general replacement strategies that we used (Step 2) are better described as categories of replacement strategies; within each category, variants of these strategies were used and investigated. Appendix B provides a comprehensive list of the strategies, along with complete statistical details. For brevity, we focus this report on a subset of these strategies. The focal strategies that we present in this report are summarized in Table 3.2.

We generate regression-based replacement scores using two approaches that primarily differ on one key dimension: accounting for idiosyncratic differences across schools. The first approach uses a straightforward OLS-based specification. Conceptually, this approach estimates a statistical relationship between prior achievement and current achievement, and we use this relationship to make out-of-sample predictions for students' achievement in 2019. This approach estimates an average relationship between current and prior achievement across all students, schools, and districts in our sample. We use different OLS models. The first OLS approach estimates an average relationship between current and prior achievement across all students, schools, and districts in a given state. It does not directly account for idiosyncratic differences across schools and districts. The second OLS approach estimates this relationship separately for each district within a given state. The benefit of the latter approach is that it allows the relationship between prior achievement and current achievement to vary across districts (e.g., different districts have different growth rates). The trade-off is that the predictions will be noisier for smaller districts. Along with the two types of OLS models that we use, we use a more advanced statistical model (i.e., a random-effects model) that directly accounts for idiosyncratic school differences while also estimating the relationship between prior and current achievement.

In Step 3, we use a simulation-based method to appraise classification consistency (Douglas and Mislevy, 2010). The objective of this analysis is to determine whether course placement recommendations based on spring 2019 MAP scores (*actual* scores) are consistent with course placement recommendations based on their replacement scores.

In Step 4 of this analysis, in which recommendations based on actual scores are compared with recommendations based on replacement scores, it is possible not only to determine how often these statuses agree exactly (which we interpret as classification consistency) but also to

Table 3.2
List of Replacement Strategies

Replacement Strategy	Description
Simple replacement	Replace spring 2019 score with spring 2018 score
Regression-based replacement	OLS prior achievement, run at state level
	OLS prior achievement, school prior achievement, run at state level
	OLS prior achievement, run at district level
	OLS prior achievement, school prior achievement, run at district level
	Random-effects prior achievement (group-mean centered), run at state level
	Random-effects prior achievement (group-mean centered), school prior achievement, run at state level
Multiple replacement	Use multiple imputation to fill in spring scores using all previous scores and demographics

investigate how often there are false positives (i.e., actual scores that would not result in recommendations for higher course placement, but replacement scores that do) and how often there are false negatives (i.e., actual scores that would result in recommendations, but replacement scores that do not).

False positives and false negatives might be particularly important in the context of COVID-19. Under regular conditions, it is known that misplacement of students into courses can have negative consequences; as we noted earlier, students who are placed into courses in which they are unable to keep pace and students who have weaknesses in key foundational areas frequently face significant difficulties reaching proficiency later on (Finkelstein et al., 2012). If there are differential or drastic effects of COVID-19 interruptions on students' prerequisite skills, misplacement might be more pronounced as a result.

For systems that use multiple measures to inform placement decisions, end-of-year assessments are not the only source of information for making placement recommendations. Therefore, there is not a one-to-one correspondence between the exact-agreement, false-positive, and false-negative rates examined in this report and what would be observed in practice. However, these analyses are useful for understanding the extent to which the information that is based on replacement scores should be weighted as though the replacement scores were actual scores. Also, we consider these analyses as providing information about consistency rather than accuracy, because the actual scores are themselves fallible. Specifically, it might be the case that, although they are used widely for such purposes, end-of-year assessment scores themselves might result in errant course placements.

In addition to examining the congruence between recommendations, we examine the extent to which replacement scores systematically overestimate or underestimate actual scores and the variability of these replacement scores. All of this diagnostic information is tabulated for the three replacement strategies, to compare and contrast them (Table 3.3). As mentioned earlier, each of these strategies was implemented within subject and grade level, and we

Table 3.3
Measures for Comparing and Contrasting Replacement Methods

Measure	Description	Interpretation
	Course placement consistency	
Exact agreement	Recommendation based on actual score agrees with recommendation based on replacement score	Percentage of students whose course placement recommendations based on test scores are consistent using replacement scores
False positive	Student is not recommended based on actual score but is recommended based on replacement score	Percentage of students whose replacement test scores indicate higher-level course placement when they should not have
False negative	Student is recommended based on actual score but is not recommended based on replacement score	Percentage of students whose replacement test scores do not indicate higher-level course placement when they should have
	Score agreement	
Bias	Average difference between replacement score and actual score	Extent to which, on average, replacement scores overestimate or underestimate actual scores
Mean square error (MSE)	Mean squared difference between replacement score and actual score	Degree of dispersion among replacement scores relative to actual scores

also estimate course placement consistency and score agreement within grade level. Complete details on these methods are available in Appendix B.

Research Question 2. Extent to Which Student and School Characteristics Influence the Consistency of Using Older Data in Course Placement Strategies

As described previously, educators and policymakers have recognized that course placement processes reliant on subjective information, such as teacher recommendations, historically have exacerbated educational disparities between black and Hispanic students and their white peers. This makes it important for schools and districts to understand whether course placement processes that employ replacement scores work equally well for all students, regardless of student or school demographics. To examine the extent to which differences in student and school demographics can influence classification consistency or differentially induce bias in replacement scores, we take the results of these analyses as outcome variables in a series of regression models in which student characteristics (race and ethnicity and gender) and school characteristics (school poverty and urbanicity) are taken as predictors.[1] These analyses allow us to examine the equity implications for each of the replacement strategies and to appraise the extent to which certain replacement methods might systematically advantage or disadvantage student subgroups or certain kinds of schools.

Study Limitations

There are several limitations to our study approach. First, we investigate a specific use case that represents only a small piece of how schools and districts make use of student assessment data. However, we believe that this particular use case is important for two reasons. First, high-functioning course placement systems play a critical role in promoting school engagement, promoting equity, mitigating failure, and reducing student dropout (Conchas, 2001; Domina, 2014; Finkelstein et al., 2012). As a result, scrutinizing course placement strategies and ensuring that there are minimal adverse implications for strategy adoption is important for advancing the work of schools. Second, although most districts have already enacted strategies regarding course placement for the current school year, there is increasing evidence that the missing student assessment data due to school closures in spring 2020 is not a one-off event and that school districts will need to continue to make decisions in ways that deviate from prepandemic practices. The impacts of COVID-19 on schools in the 2020–2021 school year are still evolving, and it is possible that fall, winter, and spring assessments will be altered or invalidated, creating a variety of novel scenarios for missing assessment data going forward. For example, California recently approved the use of shorter standardized tests in spring 2021 (Johnson, 2020).

A second limitation of this study is that it is unable to address all school assessment strategies, including the various assessments used across the country. We acknowledge that schools' and districts' needs and assessment contexts vary, and thus so will the specific strategies that are used to address missing assessments. Schools and districts use a variety of assessments, including both annual state and district assessments, such as those required by ESSA (for example, the Smarter Balanced assessments), and commercially produced tests—including interim and benchmark assessments—administered at multiple points throughout the school year (for example, the NWEA MAP Growth assessment, Renaissance Math, ACT Aspire,

and Mathematics Assessment Resource Service assessments). The assessments themselves vary widely in their reliability and level of detail. For example, in California, among districts with a placement policy, districts reported using a mix of annual state or large-scale assessments and locally developed tests (Gao and Adan, 2016). It is possible that the replacement strategies that we study using a widely used interim assessment may operate differently when applied to a state summative test or to a different commercial interim test. Rather than focus on offering a comprehensive list of situation-specific recommendations, we focus on providing an investigative framework for schools and districts that can guide and inform local decisionmaking.

The analytic strategy of this report assumes that the missing score on which a course placement decision is based is the next in a series of previously observed sequential scores, such as from annual state assessments or thrice-annual MAP assessments. When a stand-alone, one-off assessment is used for placement—and provides the missing score—two additional limitations arise. First, the simple replacement strategy does not apply, because there is no prior score to use for the imputation. Second, the extent to which our findings for the regression-based and multiple replacement strategies apply to a missing stand-alone placement assessment will depend on the strength of correlation between the stand-alone outcome and the sequence of previously observed scores. The use of a separate stand-alone assessment for placement would suggest that either the correlation with the standard assessment is not overly strong or the precision of the standard assessment at the cut point for assessment is weak. Although the regression-based and multiple replacement strategies are applicable for a missing stand-alone placement assessment, evaluation of the quality of these strategies in this case is beyond the scope of this study.

Finally, although the investigations considered in this report help schools and districts by characterizing the relative strengths and weaknesses of several score-replacement strategies and illustrating the intended and unintended consequences for the consistency and equity of course placement processes, all of these analyses were conducted on data collected under normal operating conditions (pre-COVID-19). Although fall 2020 assessment data are available for some districts and potentially could be used to shed important light on the unique ways in which the pandemic may have adversely affected these decisionmaking processes, many districts opted not to conduct academic assessments in fall 2020. Even in districts that did administer assessments, not all students participated in the testing process. One critical issue is that there appear to be systematic differences between students who participated in fall 2020 assessments and those who did not: Specifically, black and Hispanic students, students with lower prior achievement, and students in schools with higher concentrations of economically disadvantaged students were less likely to participate in fall 2020 assessments (Johnson and Kuhfeld, 2020). We believe that analyses of fall 2020 data would be compromised by these differences in the test-taking population. In particular, the students who were most likely to be adversely affected by the pandemic—and the students for whom the equity of course placement processes is a paramount concern—are the same students who are most likely to be missing from our fall 2020 MAP data.

CHAPTER FOUR

Results

In this chapter, we first present results about the characteristics of the replacement scores created using three different strategies (simple replacement, regression-based replacement, and multiple replacement), focusing on how each particular strategy influences the consistency of course placement recommendations based on test scores. We then present the results of our explorations of the extent to which student, school, and district characteristics are associated with differences in replacement score quality. Although we conducted our analyses separately by grade level, our analyses suggested that the results did not differ substantially by grade, and so, to simplify our exposition, we summarize all of our findings across grade levels in this report. Complete grade-level results are available from the authors upon request.

Using Older Assessment Information to Replace Missing Test Scores in Course Placement Processes

In this chapter, we summarize our results. Complete details, including standard deviations, minima, and maxima, are available in Appendix C. Figure 4.1 displays the five measures that we used to compare and contrast our replacement methods. The "Bias" and "MSE" columns summarize information about the extent to which the replacement scores agree with the actual scores. Bias provides an estimation of the extent to which, on average, the replacement scores overestimate or underestimate the actual scores. Underestimation is indicated by red bars, while overestimation is indicated by green bars. MSE provides an estimate of the degree of dispersion among the replacement scores relative to the actual scores. It is generally considered desirable to have small MSE and to have bias that is close to zero.

The three columns on the right summarize information about the extent to which course placement recommendations based on replacement scores are consistent with course placement recommendations based on actual scores. The exact-agreement column ("Exact") gives the percentage of students whose course placement recommendations based on actual scores are consistent using replacement scores. The false-positives column ("FP") gives the percentage of students whose replacement test scores indicated a higher-level course placement when they should not have. The false-negatives column ("FN") gives the percentage of students whose replacement test scores did not indicate higher-level course placement when they should have. In general, it is desirable to have exact-agreement rates close to 100 percent, which would indicate that the recommendations based on replacement scores are completely consistent with the recommendations based on actual scores. We refer to Figure 4.1 as we highlight several themes that emerge across the results.

Figure 4.1
Comparison of Replacement Strategy Bias and Consistency

Model	Bias	MSE	Exact	FP	FN
Math					
Simple replacement					
Replace spring 2019 score with spring 2018 score	–0.01	0.3	84%	8%	8%
Regression-based replacement					
OLS (state level)	–0.01	0.4	79%	8%	14%
OLS with school prior achievement (state level)	–0.01	0.2	85%	6%	9%
OLS (district level)	–0.01	0.3	82%	7%	11%
OLS with school prior achievement (district level)	0.02	35.5	84%	7%	9%
Random effects (state level)	–0.13	0.4	78%	5%	17%
Random effects with school prior achievement (state level)	–0.02	0.2	85%	6%	9%
Multiple replacement					
Multiple imputation	–0.01	0.2	84%	7%	8%
Reading					
	Bias	MSE	Exact	FP	FN
Simple replacement					
Replace spring 2019 score with spring 2018 score	0.03	0.3	82%	10%	9%
Regression-based replacement					
OLS (state level)	0.01	0.4	78%	8%	14%
OLS with school prior achievement (state level)	0.01	0.3	83%	7%	10%
OLS (district level)	0.02	0.3	80%	8%	12%
OLS with school prior achievement (district level)	0.03	18.8	82%	8%	10%
Random effects (state level)	–0.08	0.4	77%	6%	17%
Random effects with school prior achievement (state level)	0.00	0.3	83%	7%	11%
Multiple replacement					
Multiple imputation	0.02	0.3	82%	9%	9%

NOTE: FN = false negative; FP = false positive. Values displayed are weighted averages across districts, weighted by district size.

Consistent Course Placement Decisions Can Be Made Using All Three of the Replacement Strategies, Although Much Depends on the District Context

All three classes of replacement strategies that we investigated were capable of producing consistent course recommendations. On average, the exact-agreement rates were between 78 percent and 85 percent for all of the strategies that we considered (Figure 4.1). The most-consistent course placement recommendations were based on regression-based methods that used information from all tested students in a state and that incorporated school prior achievement into the regression model. However, both the simple replacement and multiple replacement methods produced consistent recommendations, on average, for both subjects, with exact-agreement rates between 82 percent and 84 percent.

The least-consistent course placement recommendations were based on the regression-based methods that do not use school-level prior achievement in the prediction models. In fact, these replacement methods were the only methods we tested in which the exact-agreement rates were less than 80 percent. Exact agreement is lower for these methods because they systematically generate projected scores that are lower than students' actual scores. This corre-

sponds to lower exact agreement and, in particular, to having more students not recommended for higher-level course placement on the basis of replacement scores who would have been recommended for higher-level placement using actual scores.

However, although all of our tested replacement strategies proved capable of producing consistent course placement recommendations on average across all students and districts, this does not mean that the replacement strategy selected is of no consequence. In fact, there are important district-to-district differences in how well these strategies perform, and the overall exact-agreement rates do not describe all districts equally well. For example, although the simple replacement methods had average exact-agreement rates that were around 84 percent, there were some districts with 100-percent exact agreement and other districts with exact agreement as low as 62 percent. For the regression-based replacement methods, there were some districts with exact-agreement rates as low as 9 percent, suggesting that regression-based strategies could result in nine of every ten students being misidentified for course placement based on replacement test score values.

There are several explanations for these district-to-district differences. For the simple replacement strategy, the district with the least-consistent placement recommendations had a significant proportion of students who either underperformed or overperformed their prior-year scores. In other words, about one-third of the tested students in the district had actual scores that were either significantly higher or significantly lower than their prior-year scores. Less consistency might also be observed in districts where there is a significant number of students whose actual scores are very close to the cutoff for course recommendations. In this case, students are likely to end up with different recommendations if their replacement score differs even slightly from their actual score. For the regression-based replacement strategy, the districts with the least-consistent placement recommendations had a smaller number of tested students. Any given projection can be thrown off by idiosyncratic differences among students, but these tend to average out as the sample size increases. In fact, on average, as the number of tested students in a district increases, so does the quality of the projections. Almost all of the districts with at least 250 test score observations had exact agreement well above 50 percent, and the districts with minimum agreement rates (9 percent) had fewer than 50 tested students.

Regression-Based Methods Are More Problematic When the Regressions Are Conducted Within Districts

As we have noted, the regression-based replacement methods can be estimated within each district or across all districts in a given state. Although some states (e.g., Ohio, Pennsylvania, and Tennessee) make student score predictions available to districts, in practice, districts might not have access to the data necessary to conduct statewide analyses. This means that the regression-based methods that can be enacted within districts potentially play an important role.

On the one hand, we found that district-based results using regression-based replacement scores do not differ substantially, on average, from the actual scores, and around 80 percent of the course placement recommendations based on these replacement scores agree with the recommendations based on the actual scores. In fact, the differences between the replacement scores and the actual scores tend to be less than 1 percent of a standard deviation in both math and reading. Although it appears that replacement scores in math slightly underestimate actual scores and replacement scores in reading slightly overestimate actual scores, in practical terms, these differences are small and might simply reflect small differences in the analytic sample used in each analysis.

On the other hand, it is questionable whether regression-based methods can be enacted in all districts, and there is some evidence from these analyses that, for small districts in particular (districts with fewer than four schools), using regression-based methods can induce a large amount of uncertainty into the prediction model. In particular, once school characteristics are added into the prediction model, the replacement scores become highly volatile, and some of the replacement scores are wildly implausible. The results for districts with just a few schools are highly sensitive to model specification (e.g., whether and how the school-level variables enter the model) and can lead to heavily biased predictions and unreasonably large average errors. In examining these patterns, we did not find a simple one-size-fits-all solution, further reducing the utility of this approach for small districts.

As a matter of policy, both Pennsylvania and Tennessee, for example, do not provide model estimates for schools with fewer than ten tested students in a given grade, subject, or course (SAS, 2017). Overall, this instability seriously compromises the utility of regression-based methods that are applied within districts, and such methods are most likely useful only for districts that have a large number of students and schools.

The Assumption of Average School Experiences Might Be Problematic for Course Placement Decisions Based on Regression-Based Methods

As mentioned earlier, the regression-based replacement method uses school random effects but does not incorporate school-level prior achievement results in the least-consistent decisions. Such results can be taken as a test of the underlying assumption that students encounter average schooling experiences. Because school quality varies considerably, such an assumption could have the effect of systematically overestimating some students' future achievement and underestimating other students' future achievement. For the purpose of making individual course placement decisions, this assumption about average school quality might be particularly problematic. However, incorporating school-level information into the prediction model mitigates this issue almost entirely.

Extent to Which Student and School Characteristics Influence the Consistency of Using Older Data in Course Placement Strategies

In this section, we present our results from our examinations of the equity implications for each of the replacement strategies to appraise the extent to which certain replacement methods might systematically advantage or disadvantage different student subgroups or certain kinds of schools. We highlight two themes that emerged from our investigations of the extent to which replacement score quality is associated with student and school characteristics. Because the general patterns of results are consistent across math and reading test scores, we present the math results in this section for brevity. Complete tables of regression results, including parameter estimates and standard errors, are available in Appendix C.

There Is Evidence of Differential Method Performance Based on Student Race and Ethnicity

Figure 4.2 shows the results from our regressions that investigated the extent to which the quality of the replacement scores differs systematically by student race and ethnicity. The full models (available in Appendix B) incorporate both race and gender covariates. In the figure, we present the bias separately for black, Asian, Hispanic, and white subgroups for comparison.

Figure 4.2
Comparison of Replacement Strategies by Student Race and Ethnicity

Math						
		Bias	MSE	Exact	FP	FN
Replace spring 2019 score with spring 2018 score	White	0.0	0.4	79%	9%	12%
	Black	0.0	0.4	84%	7%	9%
	Asian	–0.1	0.4	83%	6%	12%
	Hispanic	0.0	0.4	82%	7%	11%
OLS (state level)	White	–0.1	0.4	75%	6%	20%
	Black	0.3	0.6	79%	11%	9%
	Asian	–0.3	0.6	72%	4%	25%
	Hispanic	0.2	0.4	79%	9%	11%
OLS with school prior achievement (state level)	White	0.0	0.3	79%	6%	15%
	Black	0.0	0.4	84%	4%	11%
	Asian	–0.2	0.4	82%	4%	14%
	Hispanic	0.0	0.3	82%	5%	14%
Random effects (state level)	White	–0.2	0.5	72%	5%	23%
	Black	0.2	0.5	82%	8%	10%
	Asian	–0.4	0.7	69%	3%	28%
	Hispanic	0.1	0.4	80%	8%	12%
Random effects with school prior achievement (state level)	White	–0.1	0.3	78%	6%	15%
	Black	0.0	0.4	84%	4%	12%
	Asian	–0.2	0.4	82%	4%	14%
	Hispanic	0.0	0.3	82%	4%	14%
Multiple imputation	White	0.0	0.3	80%	11%	10%
	Black	0.0	0.3	86%	7%	7%
	Asian	0.0	0.3	85%	8%	8%
	Hispanic	0.0	0.3	83%	8%	8%

NOTE: FN = false negative; FP = false positive. Results for bias and MSE are based on linear regressions. Results for exact agreement, false positives, and false negatives are based on logistic regressions and are reported as percent probabilities.

For the exact agreement, false positives, and false negatives, the percentages can be interpreted as the percentage chance for each subgroup. For example, an exact-agreement percentage of 75 percent for black students would indicate that we would expect course placements to agree for three out of every four black students, on average.

In this figure, a few things are noteworthy. First, the simple replacement and multiple replacement methods do not seem to induce differential bias based on student race and ethnicity, although the simple replacement method slightly underestimates scores for Asian students. The bias estimates (in the "Bias" column) are close to zero for white, black, and Hispanic students. However, the regression-based methods do induce differential bias based on student demographics. In particular, the regression models that do not include school prior achievement systematically underestimate scores for white and Asian students and systematically overestimate scores for black and Hispanic students. We interpret this result as providing additional evidence not only that the assumption of an average school experience is not tenable overall but also that this assumption might be differentially problematic for particular student subgroups. This interpretation reflects a long history of research showing that persistent school segregation often means that black and Hispanic students attend schools with less

qualified teachers and fewer resources compared with their white peers (Darling-Hammond, 2004; Knoeppel, 2007; Lankford, Loeb, and Wyckoff, 2002; Oakes, 1985).

The story with regards to the consistency of course placement recommendations is slightly more nuanced. First, we can see from Figure 4.2 that, across all methods, black and Hispanic students have a higher overall percentage chance of exact agreement than their white peers. For example, when the simple replacement method is used, the percentage chance of an exact agreement between the recommendation based on a replacement score and the recommendation based on an actual score is 84 percent for black students, 82 percent for Hispanic students, and 79 percent for white students. This can be explained in part by the fact that fewer black and Hispanic students are close to the 65th percentile of the score distribution; in general, the farther an observed test score is from the cut score, the less likely it is to result in a contradictory course placement recommendation.

It is worth noting, however, that the bias that we observe in the replacement scores using the regression-based replacement methods has implications for the nature of the inconsistencies in course placement recommendations. For white students, the underestimation of scores results in higher false-negative rates. When the random-effects model is used without accounting for school prior achievement, there is almost a one-in-four chance that a white student will not be recommended for course placement based on that student's replacement test score, even when the student's actual test score would result in a recommendation. By contrast, for black and Hispanic students, the fact that the replacement scores overestimate the actual scores results in higher false-positive rates; the probability that black and Hispanic students will be recommended for course placement based on replacement scores (when their actual scores suggest that they will not be recommended) is between 3 percent and 5 percent higher than for white students. The regression-based replacement methods that account for school prior achievement have much lower bias than those that do not; this is because school prior achievement largely controls for differences in school quality and mitigates the issues that are caused by assuming that all students encounter average schooling experiences. However, the use of such variables raises other important considerations for schools and districts that are considering using these replacement methods. Because of the school segregation issues described earlier, and because black and Hispanic students often attend lower-quality schools because of factors largely outside their control, including school prior achievement in the regression-based replacement models can send a message to stakeholders that there are low achievement expectations for students based solely on race and ethnicity.

There Is Evidence of Differential Method Performance Based on School Poverty

At a school level, we see similar patterns based on school poverty, which we measure using the proportion of students that are eligible for FRPL. In Figure 4.3, we display information about score consistency for three levels of school poverty: 0 percent of FRPL-eligible students, 50 percent of FRPL-eligible students, and 100 percent of FRPL-eligible students. The simple replacement and multiple replacement models are not affected by school poverty, and, when these models are used, the replacement scores are, on average, very similar to the actual scores, regardless of school poverty context. However, the regression-based methods are sensitive to school poverty. In particular, these methods tend to overestimate scores for students in high-poverty contexts and underestimate scores for their more advantaged peers. Again, this is not surprising; poverty is strongly associated with school quality, teacher quality, and access to resources (Lankford, Loeb, and Wyckoff, 2002; Reardon, 2019). This again shows how assum-

Figure 4.3
Comparison of Replacement Strategies by School Poverty

		Math				
		Bias	MSE	Exact	FP	FN
Replace spring 2019 score with spring 2018 score	0% FRPL	0.0	0.2	82%	9%	9%
	50% FRPL	0.0	0.3	85%	7%	8%
	100% FRPL	0.0	0.3	87%	6%	7%
OLS (state level)	0% FRPL	–0.4	0.5	71%	2%	29%
	50% FRPL	0.1	0.5	77%	9%	16%
	100% FRPL	0.6	0.6	82%	16%	3%
OLS with school prior achievement (state level)	0% FRPL	0.0	0.2	82%	9%	10%
	50% FRPL	0.0	0.3	85%	6%	10%
	100% FRPL	0.0	0.3	87%	4%	10%
Random effects (state level)	0% FRPL	–0.5	0.5	67%	0%	33%
	50% FRPL	0.0	0.5	76%	6%	18%
	100% FRPL	0.5	0.5	86%	12%	4%
Random effects with school prior achievement (state level)	0% FRPL	0.0	0.2	82%	8%	11%
	50% FRPL	0.0	0.3	85%	6%	11%
	100% FRPL	0.0	0.3	87%	4%	10%
Multiple imputation	0% FRPL	0.0	0.2	83%	9%	9%
	50% FRPL	0.0	0.2	86%	7%	8%
	100% FRPL	0.0	0.3	88%	5%	7%

NOTE: FN = false negative; FP = false positive. Results are based on linear regressions conducted at the school level.

ing average school experiences, by not accounting for systematic differences in the kinds of school experiences to which students have access, has drastically differential impacts on the consistency of student placement recommendations. For example, when the random-effects method is used, the downward bias results in much higher false negatives for more-advantaged schools. The false-negative rate for schools serving no FRPL-eligible students is 33 percent: Nearly one in three students who would be recommended for course placement based on actual scores are not recommended based on replacement scores. For schools serving high-poverty communities (100-percent FRPL-eligible), the false-positive rate is 12 percent.

Although we considered these issues in separate analyses, it is important to note that socioeconomic status and race are not independent. Our results do not remove, for example, the influence of socioeconomic status on the relationship between race and our replacement strategies. The correlation of these two factors, and the influence of race and racism on the socioeconomic dynamics in the United States, make it difficult to isolate the independent effects of these characteristics on the consistency of replacement strategies. For this reason, we prefer to present our results in separate analyses.

Summary and Discussion

Many schools and districts rely on test scores to guide course placement processes. Sometimes, school districts look for evidence that students have exceeded grade-level standards to support placing students into accelerated courses of study; other times, school districts might look for evidence that a student's achievement in a subject compares favorably with that of other students in a reference population to recommend that student for course placement. Although test scores are not (and should not be) the only factor guiding course placement decisions, they

play a critical role in ensuring that all students, particularly underrepresented students, have fair access to advanced coursework and equitable opportunities to learn. This is particularly true given that there are persistent concerns about racial biases in course placement processes that rely heavily on more-subjective criteria, including teacher recommendations (Gamoran, 1992; Kelly, 2007).

However, course placement processes for the 2020–2021 school year were largely upended by the pandemic, because information about student achievement and what students know and can do with respect to grade-level standards is typically derived from end-of-year tests. This year, because of widespread school closures and stoppages to state assessment programs, such information was not available to schools and districts as input into these decision processes, and there was a great deal of ambiguity about how to select students for specialized programming and accurately place students into courses.

In this report, we investigated three strategies that schools and districts could use to ameliorate the problem of missing spring 2020 assessment data: (1) simple replacement, (2) regression-based replacement, and (3) multiple replacement. We investigated whether these replacement methods could lead to course placement recommendations that were consistent with recommendations that would be made based on actual test scores. We also investigated whether these replacement methods had any unintended consequences for equity, focusing on the extent to which these replacement strategies could differentially advantage or disadvantage student subgroups or certain kinds of schools. Although our results are situated in the context of course placement, they would, in principle, be applicable to any criterion-referenced decisions made by schools or districts. Additionally, the investigative methods that we used in this report—comparing and contrasting the replacement scores with actual scores and documenting bias, variability, and consistency—can be readily applied to other use cases.

The strengths and areas that merit consideration by schools and districts for these replacement strategies are summarized in Table 4.1. We highlight several of these considerations in the remainder of this chapter.

Although there was evidence that all of these replacement strategies could yield consistent course recommendations, there was considerable variability in how well all of the replacement strategies performed across districts. In some districts, misclassification rates were quite high. Districts should be aware that our analyses show that regression-based replacement methods improve as the number of tested students increases and that the districts with the least-consistent placement recommendations had fewer than 250 tested students. Districts should also be aware that simple replacement strategies might result in high rates of misclassification for students who have significant changes in achievement or content mastery between testing periods. In addition, students whose actual scores are just above or just below the criterion for course recommendations are likely to end up with different recommendations if their replacement scores differ even slightly from their actual scores. In other words, it is possible to have a replacement strategy that is unbiased on average, but, given sufficient prediction uncertainty, it still might yield relatively inaccurate decisions, particularly for students who are near the cutoff.

This strongly suggests that schools and districts should not rely solely on replacement scores to inform placement decisions (consistent with current best practices that rely on multiple measures) and that districts should continue to monitor student course performance after placement to provide appropriate supports, adjust course placements, and ensure student success. There are significant questions about the tenability of the assumption of average school

Table 4.1
Strengths and Considerations for Replacement Strategies

Replacement Strategy	Strengths	Considerations
Simple replacement	• Provides information that is generally consistent with actual scores, and there is little evidence that consistent course placement recommendations are affected by student demographics or school characteristics • Can be readily implemented by schools and districts	• Might not perform well in districts where subgroups of students have significant changes in achievement or content mastery between testing periods • Relies on an assumption that the replacement test is measuring the same content and that the replacement test scores can be interpreted in the same way as the actual test scores • Limited use when there are not contiguous assessment results
Regression-based replacement	• These models are similar in nature to growth models or value-added models used in many state and district accountability systems • Student projections can inform placements for future grades and courses, which might be helpful for high school placement processes and courses for which prior-year tests are not available	• When school prior achievement is not accounted for, might result in overestimation of scores for minority students or students who attend high-poverty schools and underestimation of scores for majority students or those who attend low-poverty schools • Requires statistical software and experience with data analysis to implement, and might require access to statewide data • With fewer than 250 tested students, decisions based on replacement scores might not be consistent
Multiple replacement	• Provides information that is generally consistent with actual scores, and there is little evidence that course placement recommendations are affected by student demographics or school characteristics	• Complex implementation that does not result in decisions that are noticeably more consistent than when simpler, more-straightforward methods are used

quality for the regression-based replacement strategy. In particular, because regression-based replacement models do not recognize that black, Hispanic, and economically disadvantaged communities systematically have access to lower-quality schools than their white, Asian, and more advantaged peers, these models overestimate test scores for black and Hispanic students or students who attend high-poverty schools and underestimate test scores for white students, Asian students, and students who attend more–economically advantaged schools. On average, this results in more student recommendations for advanced course placement than would occur on the basis of actual test scores for black, Hispanic, and economically disadvantaged students and fewer recommendations than would occur on the basis of actual test scores for their white, Asian, and more advantaged peers.

Taken together, these two findings suggest that the degree to which any replacement methods work is a function of school and district context. In particular, by design, the regression-based replacement methods make predictions that do not depend on students' educational context; however, this particular use (course placement processes) is, by its very nature, embedded in a specific context, and this lack of alignment poses problems for bias and consistency. Although as a rule such properties should be considered undesirable features of a score-replacement method, our results suggest that regression-based replacement methods tend to overestimate the performance of black and Hispanic students and students who attend schools with higher concentrations of economically disadvantaged students. If administrators

and policymakers want to err on the side of being cautious for historically disadvantaged students of color or schools operating in high-poverty contexts, the increased false-positive rates and decreased false-negative rates might, in fact, be desirable. In other words, if the goal of a school or district is to adopt course placement policies that compensate for structural and societal mechanisms that result in inequitable sorting of students into schools, score-replacement strategies that do not penalize vulnerable students near the margin of the placement cutoff might, in fact, help achieve that policy goal.

Of course, relying on overestimation as a means to equalize opportunities to learn also has potential pitfalls that districts should consider. Although failing to provide equitable access to advanced courses certainly has adverse consequences for underrepresented students, it is also the case that there are adverse consequences for students who are placed into courses for which they are not prepared, including increased rates of dropout (Finkelstein et al., 2012). Again, these findings strongly suggest that if districts are relying on replacement scores to inform course placement decisions, they should continue to monitor student course performance after placement to provide appropriate supports, adjust course placements, and ensure student success.

Finally, not all of these methods are easily implementable at a school or district level; some of the regression-based methods in particular perform best when they are implemented using statewide data. It might be that, because of staff capacity or district size, the only viable option for schools and districts is to use a simple replacement strategy. In this case, it is important for schools and districts to consider whether the replacement test is measuring the same content, and it is important that scores have the same (or at least similar) interpretations.

We caution that our results might be affected by several aspects of our research design and that schools and districts will want to interrogate these issues further to understand the extent to which our conclusions apply to course placement processes more broadly. First, we selected a cut point (the 65th percentile) that is near the center of the score distribution. We anticipate, based on prior work (e.g., Doan, Schweig, and Mihaly, 2019; Douglas and Mislevy, 2010; Martinez, Schweig, and Goldschmidt, 2016), that decision consistency might be related to the location of the cutoff and the number of cutoffs for course placement. We conducted a sensitivity analysis to appraise the extent to which our conclusions are sensitive to cut-score location, and this analysis suggests that our results can be generalized to describe the consistency of replacement strategies for cut points that are between the 45th and 75th percentiles. Although cut scores that are more extreme (for example, at the 95th percentile) might show differences in performance, such scores might be more influenced by regression to the mean and score uncertainty. The MAP scores are computer-adaptive, so even scores in the tails of the distribution are relatively precise. This feature limits the generalizability of our results at the extremes of the scoring distribution relative to other assessment scenarios that do not use adaptive algorithms.

A second aspect of our research design that might affect our results is that the NWEA MAP assessments have high test-retest reliability, meaning that the correlations of the test scores across administrations are quite high; the winter-spring correlations average around 0.86 in grades 2–8, and the spring-to-spring correlations are around 0.80. The performance of the simple replacement method in particular was most likely buoyed by these high correlations; we would anticipate that, for assessment systems with lower test-retest reliabilities, conclusions about the robustness of the simple replacement systems might change. Although past research on classification consistency suggests that the impacts of changes in correlation are minimal

when correlations are in the range of 0.5 to 1.0, correlations below this might be problematic (Wan, Brennan, and Lee, 2007). We caution against interpreting this as a reference criterion for minimum correlations. Even when district-level correlations are in this range, it is possible that, for subgroups of students, correlations might differ in magnitude and even in direction. Additionally, high correlations in and of themselves do not suggest that two test scores have equivalent interpretations.

Finally, the data that we had were available only for elementary and middle school students, and all of the methods that we considered were applied to students who had relevant testing information available in the prior grade. Although such data are able to address course placement processes for typical elementary and middle schools, these data are less well suited to address placement processes in high schools, because many high school students do not have contiguous assessment results. In particular, the simple replacement methods that we investigated rely on there being a recent relevant assessment that is a viable substitute for the missing assessment. The regression-based methods that we explored have been applied in some states to create projected student scores for courses that are not typically offered in contiguously assessed grades (for example, Advanced Placement mathematics courses), but it is a limitation of our study that we cannot fully explore the implications for high schools using the data at hand.

CHAPTER FIVE

Implications for Decisionmaking in the Absence of Large-Scale Test Scores

Our report focuses on the efficacy of score-replacement strategies under normal circumstances by contrasting three potential methods. Under these conditions, two strategies—simple replacement and regression-based replacement—depending on the intended purpose, provide a reasonable projection of how a student may have performed in the absence of an observed assessment score. The results also suggest that it is important to think about whether the assumption of a student attending an average school is important for the intended use. However, the analysis and results that we have thus far presented do not speak directly to estimating replacement scores during the period affected by the COVID-19 pandemic.

The COVID-19 pandemic has caused unprecedented disruptions to schooling in the United States and across the world. Researchers are already attempting to measure how much the pandemic has affected students' academic achievement and attainment. Although this work is early and ongoing, preliminary results suggest that students are struggling in math, on the order of 5 to 10 percentile points, and many districts are reporting higher-than-normal shares of students failing courses (Barnum, 2020; Kuhfeld, Tarasawa, et al., 2020). Furthermore, given the differential impact of COVID-19 on communities of color, black and Hispanic students likely have experienced a disproportionate impact on their learning. Unfortunately, these students also were less likely to take fall assessments, so the measurement impact on these students is less clear.

When we combine the results of our projection methods with what is known so far about the impact of COVID-19 on student learning, there are some clear implications for schools and districts as they think about the use of assessment data for course placement. First, many of the methods addressed in this report highlight that educators, policymakers, and researchers can use students' prior test history to get a reasonable idea of how students would otherwise perform in the absence of an observed test score. At a high level, one way to think about this projection during the COVID-19 era is as a forecast of students' success in a course under normal school conditions. If students are currently performing well below this projection, it is not necessarily an indication that the projection was wrong but rather that the student is performing below where he or she otherwise would have been. Even if the replacement test score does not factor heavily into school or district decisionmaking, understanding the extent to which students are not achieving or demonstrating content mastery as forecasted could provide important information to schools or districts in identifying students who are struggling with academics, social experiences, or other aspects of well-being. This is not a novel idea, but it is important to take a step back to remember what test scores represent.

Second, it is important to think about the equity implications of how assessments are used for course placement in general, but particularly for this school year and the next. Our results suggest that regression models that do not account for school context generate projections for traditionally underserved students that have a positive bias (i.e., they are higher, on average, than their observed scores). If the goal is to use the test score as a proxy for how the student would perform under normal conditions, and use it as one of many sources of data for course placements, then this bias could be potentially equity enhancing (see Ehlert et al., 2016, for a similar example in the school accountability literature). This might be especially important for students who are experiencing the full force of COVID-19's negative effects on schooling and their communities. However, if the goal is to create a score that is more of a diagnostic indicator for students who might be struggling, then it is important to account for context, especially during COVID-19.

Finally, even in an ideal setting, with many valid fall test scores, it is unlikely that any algorithm or regression-based method will be able to fully account for the differential impact of COVID-19 on students' learning. The likely best course of action in at least the 2020–2021 and 2021–2022 school years is to pair whatever test score information is available for students with local, professional judgment (including course grades and work samples) when making course placement decisions. When doing this, however, it is important to keep in mind the potential implicit and other biases that can create differential expectations for students and affect subjective placement policies (including those that use course grades).

APPENDIX A

Data Sources for This Report

We used NWEA MAP data from NWEA's Growth Research Database (NWEA, undated). In this appendix, we provide descriptive information for all tested students and schools in the database and provide details on sample inclusion and exclusion criteria.

NWEA MAP Student Characteristics

NWEA provided data on all students who took the MAP math and reading assessments for the academic years 2016–2017 through 2018–2019. Although the MAP sample fluctuates slightly from year to year, NWEA tests more than 7 million kindergarten through eighth-grade students across 17,000–19,000 U.S. public schools each school year, representing about 23–25 percent of the public schools serving students in grades K–8. In each school year, students were tested in all 50 states and the District of Columbia, although coverage within states varies greatly; only 1–3 percent of schools in some states use MAP Growth, compared with 89 percent in another state. In addition to MAP scores, NWEA also collects data on students' age, grade level, race and ethnicity, and gender. Detailed descriptive statistics for the unique students in each school year in the analytic file are included in Table A.1.

Sample demographics remain fairly consistent from school year to school year. For example, during the 2018–2019 school year, the sample was 51 percent male, 47 percent white, 17 percent black, 4 percent Asian, 19 percent Hispanic, 2 percent Native American, and 8 percent other.

NWEA MAP Student Characteristics

School-level characteristics were obtained from the 2017–2018 Common Core of Data file from the National Center for Education Statistics. A comparison of the 15,922 schools in our sample in school year 2019–2020 relative to the U.S. population of public elementary and middle schools in the same school year (73,246 schools serving grades K–8) is provided in Table A.2. Overall, our sample closely aligns to the characteristics of U.S. public schools, including the percentage of students receiving FRPL; the percentage of urban, rural, and suburban schools; and the percentage of white students. However, the NWEA sample has a slight underrepresentation of Hispanic students compared with the national population.

Table A.1
Student Characteristics, NWEA MAP Database

Demographic Characteristic	Grade K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
2016–2017 school year										
White	0.45	0.46	0.47	0.47	0.48	0.48	0.49	0.49	0.49	0.48
Black	0.20	0.19	0.18	0.17	0.17	0.16	0.16	0.16	0.16	0.17
Asian	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Hispanic	0.17	0.17	0.18	0.18	0.18	0.17	0.18	0.18	0.17	0.17
Native American	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Other race	0.09	0.08	0.09	0.08	0.08	0.08	0.08	0.08	0.09	0.08
Male	0.52	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
N	565,253	672,815	842,892	881,473	865,291	858,373	839,564	823,519	785,043	7,134,223
2017–2018 school year										
White	0.44	0.44	0.45	0.45	0.46	0.47	0.47	0.47	0.48	0.46
Black	0.18	0.18	0.17	0.17	0.17	0.17	0.16	0.16	0.16	0.17
Asian	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Hispanic	0.17	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18
Native American	0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Other race	0.10	0.10	0.10	0.10	0.09	0.09	0.09	0.09	0.09	0.09
Male	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
N	618,644	740,659	888,938	927,546	916,805	937,966	901,989	872,273	842,552	7,647,372
2018–2019 school year										
White	0.46	0.46	0.46	0.46	0.47	0.47	0.47	0.48	0.48	0.47
Black	0.19	0.18	0.17	0.17	0.17	0.17	0.17	0.16	0.16	0.17
Asian	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.04
Hispanic	0.18	0.19	0.19	0.19	0.18	0.19	0.19	0.19	0.19	0.19
Native American	0.02	0.02	0.01	0.01	0.02	0.02	0.02	0.01	0.02	0.02
Other race	0.07	0.07	0.08	0.08	0.08	0.07	0.07	0.08	0.08	0.08
Male	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
N	599,924	700,253	836,388	858,847	861,700	886,593	866,521	827,717	790,078	7,228,021
2019–2020 school year										
White	0.47	0.46	0.47	0.47	0.47	0.47	0.48	0.48	0.48	0.47
Black	0.18	0.17	0.17	0.17	0.17	0.17	0.16	0.16	0.16	0.17
Asian	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04

Table A.1—Continued

Demographic Characteristic	Grade K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
Hispanic	0.19	0.19	0.19	0.19	0.19	0.20	0.19	0.19	0.19	0.19
Native American	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.01
Other race	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
Male	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51
N	513,382	621,366	741,284	763,542	766,195	797,042	769,369	755,284	718,074	6,445,538

NOTE: *N* = number of students.

Table A.2
School Characteristics, NWEA MAP Database

	NWEA Public Schools 2019–2020		U.S. Population of Schools Serving Grades K–8	
Characteristic	***N***	**M (SD)**	***N***	**M (SD)**
Kindergarten	11,487	69.30 (40.44)	52,587	69.82 (45.49)
First grade	11,587	69.87 (39.66)	52,904	69.07 (43.21)
Second grade	11,632	70.25 (39.54)	52,940	69.31 (42.70)
Third grade	11,645	72.59 (41.18)	52,887	71.31 (44.64)
Fourth grade	11,539	74.05 (43.95)	52,562	73.09 (47.10)
Fifth grade	11,143	76.98 (53.48)	51,185	75.38 (55.71)
Sixth grade	7,404	106.63 (108.20)	36,698	103.67 (110.58)
Seventh grade	6,040	126.12 (121.99)	30,892	121.31 (129.78)
Eighth grade	5,944	126.92 (123.69)	30,655	121.82 (131.22)
Students receiving FRPL	15,918	0.50 (0.30)	73,246	0.51 (0.31)
Hispanic students	15,919	0.20 (0.24)	73,246	0.24 (0.27)
Black students	15,919	0.16 (0.24)	73,246	0.15 (0.23)
White students	15,919	0.54 (0.33)	73,246	0.52 (0.33)
Asian students	15,919	0.04 (0.07)	73,246	0.04 (0.09)
City	15,922	0.29 (0.46)	73,259	0.27 (0.45)
Suburb	15,922	0.33 (0.47)	73,259	0.33 (0.47)
Town	15,922	0.12 (0.32)	73,259	0.12 (0.33)
Rural	15,922	0.26 (0.44)	73,259	0.28 (0.45)

SOURCE: National Center for Education Statistics, 2019.

NOTE: M = mean; *N* = number of schools with observed data for a given variable; SD = standard deviation. The enrollment variables (kindergarten through eighth grade) report the number of students enrolled in each grade, and the associated counts represent the number of schools enrolling students in that grade within the sample and population.

Details on Sample Inclusion and Exclusion Criteria

To facilitate comparisons across analytic methods, we defined a common sample of student assessment data to use for all analyses presented in this report. To ensure that the results presented in this report are driven by differences in the analytic methods themselves (as opposed to differences in the student population included in each analysis), we standardized the student population included in all analyses. Specifically, we defined a set of inclusion criteria to determine which student records would be retained in a common analytic file used for all analyses presented in this report. Our criteria for inclusion are listed in this section and are essentially the minimum level of data completeness needed for a student to be retained in all analyses. Although these inclusion criteria do result in some records being dropped from the full NWEA MAP sample, our analytic sample remains large and representative of the national K–8 student population, as shown in Table A.3.

Our analytic file consists of NWEA MAP reading and math assessment data from three school years: 2016–2017, 2017–2018, and 2018–2019. For a student to be included in the analytic file, the student had to meet the following criteria:

1. have spring assessment scores in both reading and math for school years 2016–2017, 2017–2018, and 2018–2019
2. have winter 2019 reading and math assessment scores
3. have complete student-level demographic information (race and ethnicity, gender, age, and grade level)
4. be matched to a school with a valid National Center for Education Statistics identification number
5. be assigned to a school district
6. move sequentially by grade level across school years as measured by the student's 2017, 2018, and 2019 reading and math spring assessment scores.

Sequential movement by grade level (e.g., a student in third grade in 2016–2017 must be in fourth grade in 2017–2018) across school years was added to ensure that the included students followed typical grade progressions across years. However, some students took math and reading assessments in different grade levels in the same school year. As a result, students who progressed sequentially by grade level in one subject might not have progressed sequentially in the other. Because of this discrepancy, sequential movement by grade level was considered separately for each assessment subject, resulting in a separate analytic file for each assessment subject. Specifically, we created a reading analytic file that includes only those students who have sequential reading grade levels, and a math analytic file that includes only those students who have sequential math grade levels. As a result, the student population used for the math analyses is slightly different from the population used for the reading analyses; the math analytic file consisted of 2,280,224 records, while the reading analytic file consisted of 2,280,132 records. All reading analyses were conducted with the reading analytic file, while all math analyses were conducted with the math analytic file.

The analyses presented in this report either were conducted separately by district or were pooled for all students in a state. In some cases, very few students in a school, grade, or district took a math or reading assessment. To ensure a sufficient sample size, we developed minimum thresholds to be used across analyses. For district analyses, we established a minimum sample

Table A.3
Student Demographics, NWEA MAP Analytic Sample

Demographic Characteristic	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Total
				Math				
White	0.51	0.51	0.50	0.49	0.49	0.49	0.50	0.50
Black	0.16	0.16	0.15	0.15	0.15	0.15	0.15	0.15
Asian	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Hispanic	0.16	0.16	0.19	0.19	0.20	0.20	0.20	0.19
Native American	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Hawaiian/ Pacific Islander	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Multi-race	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03
Other race	0.08	0.08	0.08	0.08	0.07	0.07	0.07	0.08
Male	0.51	0.51	0.51	0.51	0.51	0.50	0.51	0.51
N	256,531	300,096	379,707	406,187	345,089	309,638	282,976	2,280,224
				Reading				
White	0.51	0.51	0.50	0.49	0.49	0.49	0.50	0.50
Black	0.16	0.16	0.15	0.15	0.15	0.15	0.15	0.15
Asian	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Hispanic	0.16	0.16	0.19	0.19	0.20	0.20	0.20	0.19
Native American	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Hawaiian/ Pacific Islander	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Multiracial	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03
Other race	0.08	0.08	0.08	0.08	0.07	0.07	0.07	0.08
Male	0.51	0.51	0.51	0.51	0.51	0.50	0.51	0.51
N	256,536	300,044	379,688	406,182	345,070	309,627	282,985	2,280,132

NOTE: *N* = number of students. Percentages are based on students' grade level in the 2018–2019 school year.

size of 30 students for each grade level (e.g., a district had to have at least 30 second-graders for the district-grade to be included). When analyses were done at the state level, a minimum threshold of 1,000 students was used.

The schools that were included in our analytic file after the inclusion rules were applied are fairly representative of the U.S. population of schools serving grades K–8 (Table A.4). Some 30 percent of the schools in our sample are urban schools, 39 percent are suburban schools, 12 percent are schools located in towns, and 19 percent are schools located in rural areas. Our analytic sample contains a slight overrepresentation of suburban schools and a slight underrepresentation of schools located in towns.

Table A.4
School Demographics, NWEA Analytic Sample

	NWEA Analytic Sample		U.S. Population of Schools Serving Grades K–8	
Characteristic	***N***	**M (SD)**	***N***	**M (SD)**
Students receiving FRPL	13,713	0.51 (0.27)	73,246	0.51 (0.31)
City	13,713	0.30 (0.46)	73,259	0.27 (0.45)
Suburb	13,713	0.39 (0.49)	73,259	0.33 (0.47)
Town	13,713	0.12 (0.33)	73,259	0.12 (0.33)
Rural	13,713	0.19 (0.39)	73,259	0.28 (0.45)

SOURCE: All school characteristics are from the 2017–2018 Common Core of Data file from the National Center for Education Statistics.
NOTE: M = mean, *N* = number of schools with observed data for a given variable, SD = standard deviation. NWEA analytic sample data are reported using the math analytic file.

APPENDIX B

Analytic Methods

This report addresses two research questions with the objective of interrogating the consequences (intended and unintended) that might arise as a result of the strategies that schools and districts adopted to enable course placement processes in the wake of COVID-19-induced disruptions to end-of-year student assessment programs:

1. To what extent can older assessment information be used to ameliorate the problem of missing test data used for course placements?
2. To what extent do student and school characteristics influence the consistency of using older data in course placement strategies?

In this appendix, we provide technical details on three major aspects of our research plan. First, we describe the approach that we used to determine the specific focal use-case scenario that serves to provide context for our investigations. We also describe our approach to identifying the three imputation methods that we compare and contrast in our empirical investigations. Second, we provide details for how these imputation methods were implemented. Third, we provide details for the statistical models that we used to investigate how differences in student and school demographics affect strategic decisionmaking.

Identification of Use Cases and Imputation Strategies

We used a three-step process to identify the use-case and candidate imputation strategies that anchor our empirical investigations. This process consisted of (1) surveys and interviews with practitioners, (2) a systematic review of current literature that is relevant to COVID-19-related school closures, and (3) the identification of specific imputation strategies for empirical investigation.

Surveys and Interviews with Practitioners

In the first step, we elicited information about the specific issues that are of concern to schools and districts at the time of writing, including the aspects of their planning that were most affected by COVID-19-related disruptions in spring 2020 and their strategies for adjusting planning activities in response to the lack of end-of-year assessment data. We collected this information using three approaches. We first conducted a brief survey about how local education agencies are confronting issues regarding day-to-day instructional uses and student-related decisions. This survey was administered to key assessment and accountability staff members

from the 100 largest local-education agencies in the country. We received survey responses from 23 different local-education agencies; 16 of these responses were complete, meaning that respondents answered all of the survey items. Second, we conducted interviews with seven individuals using a semistructured interview protocol that prompted respondents to describe their planning strategies and the impact of COVID-19 on these activities. Finally, we convened an expert panel to provide additional information about school and district strategic thinking in fall 2020. We created spreadsheets displaying key issues and plausible strategies that districts used or considered to address their planning needs. Text from interviews and survey responses was attached to specific questions to allow for direct comparison. Three members of the research team read these responses and took notes on themes. These notes were compared at project meetings, and a final set of themes was summarized.

Review of Current Literature

In the second step, we identified current research on districts' responses to COVID-19-related school closures, past research on how schools and districts dealt with interrupted assessments or missing assessment data, and recent publications on relevant statistical methodologies. We reviewed state and district responses to test score issues associated with students opting out of standardized assessments, difficulties with online assessment implementation, literature on using student growth models to project future student scores (e.g., SAS, 2017), literature on multiple imputation to fill in missing data, literature on Bayesian data augmentation, and literature on group trajectory models (Nagin, 2010; Peugh and Enders, 2004). To provide the most-relevant guidance for the current policy, social, and economic context, we included only papers published in the past 20 years. (We also excluded papers that are unavailable in the English language.) We used major online databases, including Google Scholar and the Education Resources Information Center. All literature was coded for common features (including description of the method, important assumptions, evidence of application in real-world contexts and the specifics of those contexts, and simulation conditions). Using this systematic review, we created a summary document that characterized common issues and strategies.

Identification of Specific Imputation Strategies for Empirical Investigation

We convened an expert panel to ensure that we had identified an appropriate use case, that no important literature had been omitted from our review, and that our list of specific strategies for empirical investigation was clear and complete. Importantly, the expert panel was asked to provide input on feasibility of implementation for the proposed strategies; practical, logistical, and political considerations; and affordances, constraints, and documented evidence of performance (e.g., whether the method has been applied successfully in real-world contexts, whether the method has been shown to yield valid conclusions across a wide variety of simulation conditions). The panel also was asked to identify some of the salient issues that are presented by the specific context of COVID-19-induced interruptions. Using the input from the expert panel, we identified specific conditions for empirical investigation. The final list of our specific imputation strategies for further investigation is presented below (bolded strategies are included in the main report). There were variations for each of these imputation strategies:

1. simple replacement: take student scores on most recent available assessment in place of end-of-year assessment
    a. impute spring 2019 score with winter 2019 score

   b. **impute spring 2019 score with spring 2018 score**
   c. impute spring 2019 score with most recent available score (fall 2018 or spring 2018)

2. regression-based replacement
   a. **OLS prior achievement, run at state level**
   b. OLS prior achievement + student race and gender covariates, run at state level
   c. **OLS prior achievement, school prior achievement, run at state level**
   d. OLS prior achievement race and gender, school prior achievement race and gender, run at state level
   e. **OLS prior achievement, run at district level**
   f. OLS prior achievement + student race and gender covariates, run at district level
   g. **OLS prior achievement, school prior achievement, run at district level**
   h. OLS prior achievement race and gender, school prior achievement race and gender, run at district level
   i. **school random-effects prior achievement (group-mean centered), run at state level**
   j. school random-effects prior achievement + student race and gender covariates (group-mean centered), run at state level
   k. **school random-effects prior achievement (group-mean centered), school prior achievement, run at state level**
   l. school random-effects prior achievement race and gender (group-mean centered), school prior achievement race and gender, run at state level

3. multiple replacement
   a. **multiple imputation to fill in spring scores using all previous scores and demographics.**

Investigations of the Appropriateness of Imputation Strategies in Course Placement Decisions

To investigate the factors that schools and districts should consider when determining whether a particular imputation strategy is appropriate for course placement decisions, we conducted a series of analyses on NWEA MAP data from the 2016–2017, 2017–2018, and 2018–2019 academic years. The purpose of these investigations was to compare and contrast imputation methods to the greatest extent possible. The primary framework for this investigation follows prior work in estimating consistency and accuracy of classifications based on test scores (e.g., Douglas and Mislevy, 2010; Livingston and Lewis, 1995). As described in the main report, course placement is often based (at least in part) on policies that establish cut-score criteria. In some districts, this criterion score might be as high as the 90th percentile; in other districts, it might be closer to the median. As one interview respondent noted, "For course placement, we look at a year's worth of assessment, and the highest score would flag that a student could move to a higher course. We look at scores above the 65th percentile." We used the following criterion, based on these interviews, for our investigations: Students must score above the 65th

percentile to be recommended for higher-level mathematics courses. Our empirical investigations had four steps:

1. Record a student's observed spring 2019 MAP score.
2. Using our three imputation strategies, generate a substitute 2019 MAP score.
3. Apply our criterion to the observed MAP scores and the substitute MAP scores, and make a course recommendation for the sets of observed and substitute scores.
4. Examine the congruence between recommendations based on observed and substitute scores to determine classification consistency.

In the next section, we provide statistical details for the imputation strategies.

Statistical Details for Imputation Strategies

Take the Most Recent Test Score in Place of the End-of-Year Score

For this method, a student's score is predicted based on his or her actual score in a previous administration in the same subject:

$$Y_{i,S2019} = Y_{i,W2019}, \tag{B.1}$$

where $Y_{i,S2019}$ is the outcome for student i in spring 2019 (*S2019*) and $Y_{i,W2019}$ is the outcome for student i in winter 2019 (*W2019*). In Equation B.1, the most recent score is imputed using the score from a winter 2019 test score. The right-hand side of this equation could be modified to predict the spring 2019 score using the spring 2018 score or to predict the spring 2019 score using either the spring 2018 or fall 2019 score, whichever is more recent.

Use a Linear Model to Substitute Student Scores

Linear Ordinary Least Squares Regression with One Year of Prior Achievement

For this method, a student's score is predicted using a linear OLS regression model. Such a model is estimated in two steps. First, the following model is estimated:

$$Y_{i,S2018} = \beta_0 + \beta_1 Y_{i,S2017} + \beta_2 Z_{i,S2017} + \varepsilon_i, \tag{B.2}$$

where $Y_{i,S2018}$ is the test score for student i in spring 2018, $Y_{i,S2017}$ is the same-subject test score for student i in spring 2017, and $Z_{i,S2017}$ is an out-of-subject test score for student i. ε_i is a residual (error) term. Students' prior achievement (as well as demographics in future models) is group-mean centered (e.g., demeaned by school averages). This is akin to a school fixed-effects model and uses within-school variation in the student-level covariates. This model is used to obtain estimates of β_0, β_1, and β_2, which describe the intercept, and the conditional association of prior test scores with the outcomes. These estimates are then used to make an out-of-sample prediction for spring 2019:

$$Y_{i,S2019} = \hat{\beta}_0 + \hat{\beta}_1 Y_{i,S2018} + \hat{\beta}_2 Z_{i,S2018}. \tag{B.3}$$

We estimate this model using two different approaches. One approach estimates Equation B.3 separately for each state with at least 1,000 test takers to obtain parameter estimates. The other approach estimates model parameters in Equation B.2 separately for each district, and the out-of-sample predictions are done separately by district. In this model, we restrict our analytic sample to include only districts with 30 or more students enrolled in a given grade level.

Linear Ordinary Least Squares Regression with One Year of Prior Achievement and Student Characteristics

This same two-stage framework is expanded to include student characteristics in addition to prior achievement:

$$Y_{i,S2018} = \beta_0 + \beta_1 Y_{i,S2017} + \beta_2 Z_{i,S2017} + \beta_3 Gender_i + \beta_4 Race + \varepsilon_i, \tag{B.4}$$

where $Y_{i,S2018}$, $Y_{i,S2017}$, and $Z_{i,S2017}$ are as above and *Gender* and *Race* (race and ethnicity) represent student demographic characteristics. As with Equation B.2, the student-level covariates are demeaned by school averages. This model is used to obtain estimates of β_0 through β_4. These estimates are then used to make an out-of-sample prediction for spring 2019:

$$Y_{i,S2019} = \hat{\beta}_0 + \hat{\beta}_1 Y_{i,S2018} + \hat{\beta}_2 Z_{i,S2018} + \hat{\beta}_3 Gender_i + \hat{\beta}_4 Race. \tag{B.5}$$

Linear Ordinary Least Squares Regression with One Year of Prior Achievement and School-Average Prior Achievement

We expand Equation B.2 in this iteration of the model by including school-average prior achievement (with out-of-sample predictions for spring 2019 analogous to above):

$$Y_{i,S2018} = \beta_0 + \beta_1 Y_{i,S2017} + \beta_2 Z_{i,S2017} + \beta_1 \overline{Y}_{i,S2017} + \beta_2 \overline{Z}_{i,S2017} + \varepsilon_i. \tag{B.6}$$

Linear Ordinary Least Squares Regression with One Year of Prior Achievement, School-Average Prior Achievement, Student Demographics, and School Characteristics

The model used for this analysis expands Equation B.6 (again with out-of-sample predictions for spring 2019 analogous to above):

$$\begin{aligned} Y_{i,S2018} = {} & \beta_0 + \beta_1 Y_{i,S2017} + \beta_2 Z_{i,S2017} + \beta_3 \overline{Y_{i,S2017}} + \beta_4 \overline{Z_{i,S2017}} \\ & + \beta_5 Gender_i + \beta_6 Race_i + \beta_5 \overline{Gender_i} + \beta_6 \overline{Race_i} + \varepsilon_i \end{aligned} \tag{B.7}$$

Random-Effects Regression with One Year of Prior Achievement

For this method, a student's score is predicted using a hierarchical linear model with a school random effect:

$$Y_{ij,S2018} = \beta_0 + \beta_1 Y_{ij,S2017} + \beta_2 Z_{ij,S2017} + \alpha_j + \varepsilon_{ij}, \tag{B.8}$$

where $Y_{ij,S2018}$ is the outcome for student i in school j in spring 2018, $Y_{ij,S2017}$ is the same-subject test score for student i in spring 2017, and $Z_{ij,S2017}$ is an out-of-subject test score for student i. α_j is a school-level random effect.

This model is used to obtain estimates of β_0, β_1, and β_2. These estimates are then used to make an out-of-sample prediction for spring 2019:

$$Y_{i,S2019} = \hat{\beta}_0 + \hat{\beta}_1 Y_{i,S2018} + \hat{\beta}_2 Z_{i,S2018}. \quad \text{(B.9)}$$

Importantly, for the purpose of score prediction, we assume that the random effect from Equation B.8 is equal to zero. This is equivalent to imposing the assumption that students have an average school experience (SAS, 2017).

As with the OLS models, Equations B.8 and B.9 are expanded to include (1) student covariates; (2) school average prior achievement; and (3) one year of prior achievement, school-average prior achievement, student demographics, and school characteristics. Random-effects models are estimated within grade separately by states; only states with more than 1,000 tested students are included in model estimation.

Use Multiple Imputation Based on Chained Equations

In implementing this method, we treated all spring 2019 assessment scores as missing. Because the scores were missing for all test takers, the data were, by definition, missing completely at random (Schafer and Graham, 2002). We used R's mice package to generate five imputed data sets (Rubin, 1987). The imputation was based on an approach that used classification and regression trees as the conditional models for imputation. All available student- and school-level covariates were included in the imputation model. These were

- student gender
- student race
- prior test scores (all available)
- charter status
- magnet status
- school urbanicity
- proportion of students eligible for special education
- total school enrollment
- school-level race
- school-level FRPL eligibility
- school-level gifted status
- school-level limited English proficiency.

We then pooled the imputed data sets and took the mean imputed score as the substitute score for each student.

Investigations of Whether Consistent Decisionmaking Is Supported with Various Imputation Strategies

In this section, we provide details on how we examined the congruence between recommendations based on observed and substitute scores to determine classification consistency (Step 4). The objective of this analysis is to determine whether course placement recommendations based on spring 2019 MAP scores (*actual scores*) are consistent with course placement recom-

mendations based on their replacement scores. In the parlance of classical test theory (Douglas and Mislevy, 2010), we take students' observed spring 2019 MAP score as their *actual score* and their classification (i.e., their recommendation for course placement) as their *actual status*. We take their substitute score as a *replicate score* and their classification based on this score as their *replicate status*. We consider these analyses as providing information about consistency rather than accuracy, because the actual scores are themselves fallible. Actual and replicate classifications can then be compared using a contingency table (Table B.1).

Pairwise exact agreement (which we interpret as classification consistency) is calculated as *(a + d)/N*, where *a* is the number of students who are recommended using both scores, *d* is the number of students who are not recommended using both scores, and *N* is the total number of students. We then look at the proportions of students who are false negatives (those who are "placed down" based on their replicate score) and who are false positives (those who are "placed up" based on their replicate score). The fraction *b/N* gives the proportion of false negatives, and the fraction *c/N* gives the proportion of false positives. These measures are calculated within districts.

In addition to appraising classification consistency for each method, we examine properties of the imputed scores. Specifically, we estimate measures of bias and MSE for each of the imputation methods. Because the students' "true" spring 2019 scores are unobserved, we calculate bias by finding the average difference between the actual observed scores Θ_i and the replicate scores R_i over all students N in a specific district:

$$Bias = N^{-1}\sum_{i=1}^{N}\left(R_i - \Theta_i\right). \tag{B.10}$$

MSE is appraised using the following formula (Hoel, Port, and Stone, 1971):

$$MSE = N^{-1}\sum_{i=1}^{N}\left(R_i - \Theta_i\right)^2. \tag{B.11}$$

This provides a distribution of bias and MSE over districts.

Table B.1
Contingency Table of Math Course Placements Using Actual Scores and Replacement Scores

		Replacement Score		
		Recommend	Not recommend	Total
Actual Score	Recommend	a	b	g
	Not recommend	c	d	h
	Total	e	f	*N*

NOTE: e = a + c; f = b + d.

Examination of How Differences in Student and School Demographics Affect Decisionmaking

To examine the extent to which differences in student and school demographics influence classification consistency or differentially induce bias in substitute scores, we take the results of the analyses of prediction consistency as outcome variables in a series of regression models in which student characteristics (i.e., race and ethnicity and gender), school characteristics (i.e., school poverty, urbanicity, and school racial composition) and district characteristics (i.e., district size) are taken as predictors. These regression models allow us to examine the extent to which certain imputation methods may systematically advantage or disadvantage student subgroups or certain kinds of schools.

We first regress each of our student-level outcome variables ($Bias_i$, MSE_i, etc.) on indicators of students' race and ethnicity and gender. For the three indicators of classification consistency (e.g., whether the student's classification is an exact match, a false negative, or a false positive), we estimate logistic regression models with the standard errors clustered at the school level. For the student-level measure of bias and MSE, we estimate OLS regression models pooled across grades and states with the standard errors clustered at the school level:

$$Bias_i = \beta_0 + \beta_1 Gender_i + \beta_2 Race_i + \beta_3 \left(Gender_i * Race_i\right) + \varepsilon_i. \tag{B.12}$$

We then aggregate each of our student-level outcome variables to the school level and examine the relationship between each variable and a set of school characteristics:

$$\begin{aligned} Bias_j = {} & \beta_0 + \beta_1 Urban_j + \beta_2 Rural_j + \beta_3 Suburban_j \\ & + \beta_4 \%FRPL_j + \beta_2 RaceComp_j + \beta_3 Grade_j + \varepsilon_j. \end{aligned} \tag{B.13}$$

APPENDIX C

Complete Results

Tables C.1–C.20 present our complete regression results, including mean, standard deviations, minima, and maxima.

Table C.1
Exact-Agreement Rates: Mathematics

Replacement Strategy	Mean	Standard Deviation	Minimum	Maximum
Simple replacement				
1a	0.86	0.02	0.62	1.00
1b	0.84	0.03	0.62	1.00
1c	0.82	0.03	0.54	1.00
Regression-based replacement				
2a	0.79	0.05	0.24	0.97
2b	0.79	0.05	0.24	0.97
2c	0.85	0.03	0.59	1.00
2d	0.85	0.03	0.56	1.00
2e	0.82	0.04	0.09	0.99
2f	0.82	0.04	0.09	0.99
2g	0.84	0.04	0.09	0.99
2h	0.83	0.04	0.09	0.99
2i	0.78	0.07	0.33	1.00
2j	0.78	0.07	0.30	1.00
2k	0.85	0.03	0.56	1.00
2l	0.85	0.03	0.56	1.00
Multiple replacement				
3a	0.84	0.03	0.64	1.00

Table C.2
False-Positive Rates: Mathematics

Replacement Strategy	Mean	Standard Deviation	Minimum	Maximum
Simple replacement				
1a	0.07	0.02	0.00	0.29
1b	0.08	0.02	0.00	0.36
1c	0.11	0.03	0.00	0.46
Regression-based replacement				
2a	0.08	0.06	0.00	0.39
2b	0.08	0.06	0.00	0.39
2c	0.06	0.02	0.00	0.28
2d	0.06	0.02	0.00	0.36
2e	0.07	0.03	0.00	0.92
2f	0.07	0.03	0.00	0.92
2g	0.07	0.04	0.00	0.92
2h	0.08	0.04	0.00	0.92
2i	0.05	0.04	0.00	0.33
2j	0.05	0.04	0.00	0.33
2k	0.06	0.02	0.00	0.27
2l	0.06	0.02	0.00	0.35
Multiple replacement				
3a	0.07	0.02	0.00	0.28

Table C.3
False-Negative Rates: Mathematics

Replacement Strategy	Mean	Standard Deviation	Minimum	Maximum
Simple replacement				
1a	0.06	0.02	0.00	0.30
1b	0.08	0.02	0.00	0.31
1c	0.07	0.02	0.00	0.27
Regression-based replacement				
2a	0.14	0.08	0.00	0.76
2b	0.14	0.08	0.00	0.76
2c	0.09	0.03	0.00	0.38
2d	0.09	0.03	0.00	0.41
2e	0.11	0.03	0.00	0.41
2f	0.11	0.03	0.00	0.41
2g	0.09	0.03	0.00	0.42
2h	0.10	0.03	0.00	0.41
2i	0.17	0.10	0.00	0.67
2j	0.17	0.10	0.00	0.70
2k	0.09	0.03	0.00	0.41
2l	0.09	0.03	0.00	0.41
Multiple replacement				
3a	0.08	0.02	0.00	0.29

Table C.4
Replacement Score Bias: Mathematics

Replacement Strategy	Mean	Standard Deviation	Minimum	Maximum
Simple replacement				
1a	0.03	0.07	–0.54	0.43
1b	–0.01	0.06	–0.53	0.72
1c	0.07	0.09	–0.52	0.84
Regression-based replacement				
2a	0.01	0.30	–1.60	1.37
2b	0.01	0.30	–1.59	1.37
2c	0.01	0.06	–0.62	0.85
2d	0.01	0.06	–1.01	0.96
2e	0.01	0.07	–0.80	0.62
2f	0.01	0.24	–10.50	3.25
2g	–0.02	1.46	–103.00	5.18
2h	0.00	0.22	–3.07	2.16
2i	0.13	0.32	–1.57	1.44
2j	0.13	0.32	–1.56	1.44
2k	0.02	0.06	–0.61	0.88
2l	0.01	0.07	–1.04	1.00
Multiple replacement				
3a	–0.01	0.05	–0.45	0.41

Table C.5
Replacement Score Mean Square Error: Mathematics

Replacement Strategy	Mean	Standard Deviation	Minimum	Maximum
Simple replacement				
1a	0.17	0.04	0.05	0.72
1b	0.25	0.06	0.07	1.16
1c	0.32	0.07	0.09	1.28
Regression-based replacement				
2a	0.38	0.18	0.08	2.89
2b	0.38	0.18	0.08	2.87
2c	0.23	0.05	0.05	1.04
2d	0.24	0.26	0.05	12.90
2e	0.30	0.10	0.09	1.36
2f	0.61	9.80	0.10	530.00
2g	35.50	2,480.00	0.09	178,181.00
2h	1.40	17.90	0.10	344.00
2i	0.41	0.18	0.08	2.62
2j	0.41	0.18	0.08	2.61
2k	0.23	0.05	0.05	1.04
2l	0.24	0.26	0.05	12.90
Multiple replacement				
3a	0.23	0.04	0.09	0.68

Table C.6
Exact-Agreement Rates: Reading

Replacement Strategy	Mean	Standard Deviation	Minimum	Maximum
Simple replacement				
1a	0.84	0.02	0.65	0.99
1b	0.82	0.03	0.64	0.97
1c	0.80	0.03	0.53	0.98
Regression-based replacement				
2a	0.78	0.05	0.15	0.94
2b	0.78	0.05	0.12	0.95
2c	0.83	0.03	0.62	1.00
2d	0.83	0.03	0.63	0.98
2e	0.80	0.03	0.13	0.98
2f	0.80	0.03	0.13	0.99
2g	0.82	0.04	0.13	0.98
2h	0.81	0.04	0.13	0.99
2i	0.77	0.07	0.00	1.00
2j	0.77	0.06	0.00	1.00
2k	0.83	0.03	0.00	1.00
2l	0.83	0.03	0.00	1.00
Multiple replacement				
3a	0.82	0.03	0.62	0.98

Table C.7
False-Positive Rates: Reading

Replacement Strategy	Mean	Standard Deviation	Minimum	Maximum
Simple replacement				
1a	0.09	0.02	0.00	0.29
1b	0.10	0.02	0.00	0.29
1c	0.12	0.03	0.00	0.44
Regression-based replacement				
2a	0.08	0.05	0.00	0.39
2b	0.08	0.05	0.00	0.38
2c	0.07	0.02	0.00	0.32
2d	0.07	0.03	0.00	0.32
2e	0.08	0.03	0.00	0.87
2f	0.08	0.04	0.00	0.87
2g	0.08	0.04	0.00	0.87
2h	0.09	0.04	0.00	0.87
2i	0.06	0.04	0.00	0.60
2j	0.06	0.04	0.00	0.60
2k	0.07	0.02	0.00	1.00
2l	0.07	0.03	0.00	1.00
Multiple replacement				
3a	0.09	0.02	0.00	0.36

Table C.8
False-Negative Rates: Reading

Replacement Strategy	Mean	Standard Deviation	Minimum	Maximum
Simple replacement				
1a	0.07	0.02	0.00	0.23
1b	0.09	0.02	0.00	0.28
1c	0.08	0.02	0.00	0.34
Regression-based replacement				
2a	0.14	0.08	0.00	0.85
2b	0.14	0.08	0.00	0.88
2c	0.10	0.02	0.00	0.27
2d	0.10	0.02	0.00	0.31
2e	0.12	0.03	0.00	0.33
2f	0.12	0.03	0.00	0.32
2g	0.10	0.03	0.00	0.41
2h	0.11	0.03	0.00	0.43
2i	0.17	0.09	0.00	1.00
2j	0.17	0.09	0.00	1.00
2k	0.11	0.02	0.00	1.00
2l	0.10	0.02	0.00	1.00
Multiple replacement				
3a	0.09	0.02	0.00	0.26

Table C.9
Replacement Score Bias: Reading

Replacement Strategy	Mean	Standard Deviation	Minimum	Maximum
Simple replacement				
1a	0.03	0.05	–0.42	0.40
1b	0.03	0.05	–0.46	0.52
1c	0.08	0.08	–0.53	0.77
Regression-based replacement				
2a	–0.01	0.26	–1.19	1.24
2b	–0.01	0.26	–1.19	1.23
2c	–0.01	0.06	–0.51	0.44
2d	–0.01	0.06	–0.46	0.55
2e	–0.02	0.06	–0.54	0.52
2f	–0.01	0.21	–9.73	2.20
2g	–0.03	1.06	–74.70	4.12
2h	0.00	0.28	–1.02	5.68
2i	0.08	0.27	–3.86	1.74
2j	0.08	0.27	–3.86	1.74
2k	0.00	0.06	–1.59	1.63
2l	–0.01	0.06	–3.11	5.87
Multiple replacement				
3a	0.02	0.04	–0.39	0.49

Table C.10
Replacement Score Mean Square Error: Reading

Replacement Strategy	Mean	Standard Deviation	Minimum	Maximum
Simple replacement				
1a	0.24	0.05	0.09	1.38
1b	0.32	0.07	0.12	1.13
1c	0.39	0.08	0.13	1.20
Regression-based replacement				
2a	0.39	0.15	0.12	1.82
2b	0.39	0.15	0.12	1.81
2c	0.27	0.06	0.09	1.00
2d	0.27	0.07	0.10	1.37
2e	0.33	0.09	0.10	1.14
2f	0.56	8.10	0.10	452.00
2g	18.80	1,304.00	0.10	93,657.00
2h	1.08	11.30	0.10	256.00
2i	0.41	0.15	0.00	14.90
2j	0.40	0.15	0.00	14.90
2k	0.27	0.06	0.00	2.65
2l	0.28	0.18	0.00	91.80
Multiple replacement				
3a	0.28	0.05	0.10	0.75

Table C.11
Student-Level Regressions: Bias (Mathematics)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male	0.008***	0.013***	−0.002	−0.042***	0.009***	−0.042***	0.006***	−0.039***	0.009***	−0.041***	0.006***	−0.028***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Black	0.004+	0.002	0.024***	0.378***	0.323***	0.056***	−0.020***	0.391***	0.340***	0.060***	−0.016***	0.029***
	(0.002)	(0.003)	(0.003)	(0.008)	(0.008)	(0.003)	(0.003)	(0.008)	(0.009)	(0.003)	(0.003)	(0.002)
Male # black	−0.007***	−0.001	−0.010***	0.036***	0.032***	0.031***	0.026***	0.036***	0.032***	0.030***	0.026***	0.005**
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
Hispanic	0.001	−0.019***	−0.029***	0.251***	0.236***	0.020***	0.008**	0.285***	0.273***	0.024***	0.011***	0.019***
	(0.002)	(0.002)	(0.003)	(0.008)	(0.008)	(0.002)	(0.003)	(0.008)	(0.008)	(0.003)	(0.003)	(0.002)
Male # Hispanic	−0.005**	−0.004*	−0.004+	0.008***	0.005*	0.007***	0.003+	0.010***	0.006**	0.007***	0.004+	−0.003+
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Asian	−0.012***	−0.073***	−0.085***	−0.262***	−0.157***	−0.118***	0.002	−0.243***	−0.140***	−0.117***	0.005	−0.041***
	(0.003)	(0.004)	(0.005)	(0.013)	(0.014)	(0.004)	(0.004)	(0.014)	(0.014)	(0.004)	(0.004)	(0.003)
Male # Asian	−0.003	−0.010**	−0.016***	−0.002	−0.004	−0.011***	−0.013***	−0.002	−0.004	−0.011**	−0.013***	−0.005+
	(0.003)	(0.003)	(0.004)	(0.005)	(0.005)	(0.003)	(0.003)	(0.005)	(0.005)	(0.003)	(0.003)	(0.003)
Native	0.021***	−0.001	0.028**	0.310***	0.283***	0.029***	−0.013	0.303***	0.278***	0.030***	−0.014+	0.004
	(0.006)	(0.007)	(0.009)	(0.028)	(0.027)	(0.007)	(0.008)	(0.028)	(0.028)	(0.007)	(0.008)	(0.006)
Male # Native	0.003	0.007	−0.008	0.013	0.017+	0.021**	0.020**	0.013	0.016+	0.020**	0.021**	0.013*
	(0.005)	(0.007)	(0.008)	(0.010)	(0.010)	(0.007)	(0.007)	(0.010)	(0.010)	(0.007)	(0.007)	(0.006)
Hawaiian	0.024**	−0.039***	−0.027*	0.184***	0.165***	0.006	−0.016	0.183***	0.164***	0.012	−0.015	0.018+
	(0.009)	(0.011)	(0.013)	(0.023)	(0.022)	(0.010)	(0.011)	(0.023)	(0.022)	(0.010)	(0.011)	(0.009)

Table C.11—Continued

Characteristic	Model											
	1a	1b	1c	2a	2b	2c	2d	2i	2j	2k	2l	3a
Male # Hawaiian	−0.002	−0.020	−0.034+	0.011	0.018	0.007	0.009	0.013	0.022	0.006	0.009	−0.008
	(0.012)	(0.014)	(0.018)	(0.018)	(0.018)	(0.014)	(0.015)	(0.018)	(0.019)	(0.014)	(0.014)	(0.014)
Multi	0.018***	0.001	0.022***	0.110***	0.092***	0.026***	0.001	0.115***	0.099***	0.027***	0.004	0.018***
	(0.003)	(0.003)	(0.003)	(0.007)	(0.007)	(0.003)	(0.003)	(0.007)	(0.007)	(0.003)	(0.003)	(0.003)
Male # multi	−0.007*	−0.001	−0.010*	−0.006	−0.005	0.003	0.004	−0.004	−0.003	0.003	0.004	−0.008*
	(0.003)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.003)
Other	0.016***	0.001	0.000	0.106***	0.098***	0.022***	0.006	0.125***	0.118***	0.025***	0.002	0.011**
	(0.004)	(0.004)	(0.005)	(0.013)	(0.013)	(0.005)	(0.004)	(0.014)	(0.014)	(0.005)	(0.004)	(0.004)
Male # other	−0.002	−0.001	−0.009*	0.005	0.008*	0.005+	0.007*	0.005	0.008*	0.005	0.007**	−0.001
	(0.002)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
N	**2,280,166**	**2,280,166**	**2,280,166**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,515**	**2,279,504**	**2,279,515**	**2,279,504**	**2,219,066**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.12
Student-Level Regressions: Mean Square Error (Mathematics)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male	0.024***	0.038***	0.060***	0.049***	0.040***	0.040***	0.037***	0.056***	0.038***	0.041***	0.037***	0.025***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.002)	(0.001)	(0.001)	(0.003)	(0.002)	(0.001)
Black	0.029***	0.050***	0.043***	0.147***	0.115***	0.038***	0.059***	0.046***	0.026***	0.037***	0.059***	0.021***
	(0.002)	(0.003)	(0.003)	(0.007)	(0.007)	(0.004)	(0.018)	(0.007)	(0.006)	(0.004)	(0.018)	(0.002)
Male # black	0.008***	0.015***	0.009**	−0.003	0.032***	0.006+	0.005	−0.009**	0.026***	0.005	0.004	0.006***
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.010)	(0.003)	(0.003)	(0.003)	(0.010)	(0.002)
Hispanic	0.015***	0.026***	0.022***	0.034***	0.023***	0.014***	0.018***	−0.033***	−0.041***	0.014***	0.018***	0.009***
	(0.001)	(0.002)	(0.002)	(0.004)	(0.004)	(0.003)	(0.002)	(0.005)	(0.005)	(0.003)	(0.002)	(0.001)
Male # Hispanic	0.000	0.006**	0.001	−0.018***	0.006*	−0.003	0.003	−0.020***	0.005*	−0.003	0.003	−0.002
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.001)
Asian	−0.009***	0.007*	−0.005	0.211***	0.163***	0.020***	0.008*	0.259***	0.188***	0.023***	0.007*	−0.010***
	(0.001)	(0.003)	(0.004)	(0.014)	(0.013)	(0.004)	(0.003)	(0.017)	(0.016)	(0.004)	(0.003)	(0.002)
Male # Asian	−0.006**	−0.014***	−0.008*	0.025***	−0.001	0.003	−0.010**	0.021**	−0.001	0.003	−0.009**	0.002
	(0.002)	(0.003)	(0.004)	(0.006)	(0.005)	(0.004)	(0.003)	(0.007)	(0.006)	(0.004)	(0.003)	(0.002)
Native	0.019***	0.027***	0.043***	0.154***	0.137***	0.030**	0.043***	0.061***	0.052**	0.030**	0.043***	0.027***
	(0.004)	(0.006)	(0.007)	(0.022)	(0.022)	(0.010)	(0.008)	(0.017)	(0.017)	(0.010)	(0.008)	(0.005)
Male # Native	0.011+	0.023**	0.014	0.019	0.049***	0.008	0.044*	0.018	0.045***	0.044	0.044*	0.008
	(0.006)	(0.008)	(0.010)	(0.013)	(0.014)	(0.009)	(0.021)	(0.012)	(0.013)	(0.035)	(0.021)	(0.007)
Hawaiian	0.004	0.001	−0.017	0.028+	0.025	−0.004	0.023*	−0.024	−0.017	−0.005	0.022*	0.001
	(0.006)	(0.008)	(0.011)	(0.017)	(0.016)	(0.008)	(0.009)	(0.018)	(0.019)	(0.008)	(0.009)	(0.007)

Table C.12—Continued

Characteristic	Model											
	1a	1b	1c	2a	2b	2c	2d	2i	2j	2k	2l	3a
Male # Hawaiian	−0.007	0.003	0.009	−0.016	0.002	−0.010	−0.008	−0.017	0.001	−0.010	−0.007	0.002
	(0.009)	(0.013)	(0.016)	(0.018)	(0.018)	(0.011)	(0.013)	(0.018)	(0.019)	(0.011)	(0.013)	(0.011)
Multi	0.008***	0.011***	0.013***	0.059***	0.064***	0.007+	0.011***	0.025***	0.034***	0.007+	0.011***	0.004*
	(0.002)	(0.003)	(0.004)	(0.005)	(0.005)	(0.004)	(0.003)	(0.005)	(0.006)	(0.004)	(0.003)	(0.002)
Male # multi	0.005	0.005	0.001	0.006	0.017**	0.000	0.007	0.008	0.019***	0.000	0.006	0.002
	(0.003)	(0.004)	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.003)
Other	0.003	0.003	−0.001	0.001	−0.005	0.025	0.009**	−0.037***	−0.042***	0.025	0.009**	0.007***
	(0.002)	(0.003)	(0.003)	(0.008)	(0.008)	(0.018)	(0.003)	(0.009)	(0.009)	(0.018)	(0.003)	(0.002)
Male # other	0.002	0.003	0.002	−0.005	0.005	0.001	0.002	−0.006+	0.004	0.000	0.002	0.002
	(0.002)	(0.003)	(0.004)	(0.004)	(0.004)	(0.006)	(0.003)	(0.003)	(0.003)	(0.006)	(0.003)	(0.002)
N	**2,280,166**	**2,280,166**	**2,280,166**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,515**	**2,279,504**	**2,279,515**	**2,279,504**	**2,219,066**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.13
Student-Level Regressions: Exact Agreement (Mathematics)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male	0.046***	0.035***	0.040***	0.016***	0.059***	0.020***	0.034***	0.000	0.053***	0.014**	0.034***	0.053***
	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Black	0.358***	0.364***	0.341***	0.257***	0.333***	0.379***	0.368***	0.510***	0.556***	0.383***	0.369***	0.401***
	(0.013)	(0.013)	(0.012)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)	(0.015)	(0.014)	(0.014)	(0.013)
Male # black	0.000	−0.036**	−0.016	0.038***	−0.055***	0.002	−0.010	0.050***	−0.032**	0.008	−0.010	−0.002
	(0.012)	(0.012)	(0.011)	(0.010)	(0.011)	(0.012)	(0.012)	(0.010)	(0.011)	(0.012)	(0.012)	(0.012)
Hispanic	0.218***	0.223***	0.228***	0.258***	0.294***	0.242***	0.238***	0.442***	0.466***	0.244***	0.234***	0.259***
	(0.011)	(0.010)	(0.010)	(0.012)	(0.012)	(0.011)	(0.011)	(0.013)	(0.013)	(0.011)	(0.011)	(0.011)
Male # Hispanic	−0.077***	−0.082***	−0.085***	−0.028**	−0.106***	−0.074***	−0.089***	−0.031***	−0.096***	−0.069***	−0.084***	−0.079***
	(0.011)	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.011)	(0.009)	(0.009)	(0.011)	(0.011)	(0.011)
Asian	0.288***	0.259***	0.299***	−0.171***	−0.033	0.236***	0.288***	−0.224***	−0.070**	0.227***	0.292***	0.306***
	(0.019)	(0.019)	(0.018)	(0.022)	(0.022)	(0.019)	(0.019)	(0.022)	(0.023)	(0.019)	(0.019)	(0.020)
Male # Asian	0.036	0.030	0.029	0.112***	0.122***	0.010	0.025	0.119***	0.145***	0.016	0.030	0.031
	(0.023)	(0.021)	(0.020)	(0.017)	(0.018)	(0.021)	(0.021)	(0.016)	(0.016)	(0.021)	(0.021)	(0.023)
Native	0.319***	0.391***	0.318***	0.293***	0.335***	0.401***	0.376***	0.491***	0.508***	0.412***	0.378***	0.339***
	(0.036)	(0.036)	(0.033)	(0.039)	(0.041)	(0.037)	(0.037)	(0.050)	(0.052)	(0.038)	(0.037)	(0.038)
Male # Native	0.016	−0.050	−0.008	−0.028	−0.098**	−0.012	−0.061	−0.019	−0.074*	−0.005	−0.038	−0.006
	(0.038)	(0.037)	(0.035)	(0.028)	(0.030)	(0.037)	(0.037)	(0.030)	(0.030)	(0.038)	(0.037)	(0.040)
Hawaiian	0.369***	0.481***	0.389***	0.333***	0.342***	0.417***	0.374***	0.506***	0.521***	0.419***	0.381***	0.368***
	(0.074)	(0.072)	(0.064)	(0.063)	(0.063)	(0.069)	(0.069)	(0.065)	(0.069)	(0.068)	(0.069)	(0.072)

Table C.13—Continued

Characteristic	Model											
	1a	1b	1c	2a	2b	2c	2d	2i	2j	2k	2l	3a
Male # Hawaiian	−0.132	−0.133	−0.042	−0.003	−0.033	−0.113	−0.216*	0.008	−0.074	−0.116	−0.180*	−0.090
	(0.097)	(0.095)	(0.084)	(0.077)	(0.075)	(0.096)	(0.090)	(0.082)	(0.076)	(0.095)	(0.090)	(0.092)
Multi	0.122***	0.150***	0.118***	0.096***	0.091***	0.173***	0.154***	0.188***	0.177***	0.173***	0.156***	0.151***
	(0.016)	(0.015)	(0.015)	(0.015)	(0.015)	(0.016)	(0.016)	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)
Male # multi	−0.006	−0.034	−0.038+	−0.010	−0.050**	−0.030	−0.040+	−0.011	−0.045*	−0.026	−0.037+	−0.008
	(0.022)	(0.021)	(0.020)	(0.018)	(0.018)	(0.021)	(0.021)	(0.018)	(0.018)	(0.021)	(0.021)	(0.022)
Other	0.061***	0.076***	0.067***	0.135***	0.148***	0.066***	0.057***	0.218***	0.231***	0.065***	0.057***	0.067***
	(0.013)	(0.014)	(0.012)	(0.020)	(0.020)	(0.014)	(0.014)	(0.023)	(0.023)	(0.014)	(0.014)	(0.014)
Male # other	−0.024	−0.016	−0.019	−0.013	−0.035**	0.003	−0.011	−0.005	−0.027*	0.008	−0.014	−0.023
	(0.015)	(0.015)	(0.014)	(0.013)	(0.013)	(0.015)	(0.015)	(0.013)	(0.013)	(0.015)	(0.015)	(0.015)
N	**2,280,166**	**2,280,166**	**2,280,166**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,219,066**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.14
Student-Level Regressions: False Positive (Mathematics)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male	−0.022**	0.006	−0.060***	−0.094***	0.110***	−0.158***	0.049***	−0.089***	0.125***	−0.153***	0.052***	−0.129***
	(0.007)	(0.007)	(0.006)	(0.008)	(0.008)	(0.008)	(0.008)	(0.010)	(0.010)	(0.008)	(0.008)	(0.007)
Black	−0.385***	−0.352***	−0.323***	0.588***	0.418***	−0.469***	−0.813***	0.662***	0.492***	−0.451***	−0.790***	−0.418***
	(0.015)	(0.017)	(0.014)	(0.021)	(0.023)	(0.019)	(0.023)	(0.023)	(0.026)	(0.020)	(0.025)	(0.017)
Male # black	−0.032+	0.026	0.013	0.000	−0.038**	0.043*	0.031	−0.011	−0.054**	0.039*	0.033	−0.018
	(0.017)	(0.016)	(0.014)	(0.014)	(0.014)	(0.018)	(0.020)	(0.016)	(0.017)	(0.019)	(0.021)	(0.018)
Hispanic	−0.278***	−0.258***	−0.337***	0.402***	0.350***	−0.399***	−0.466***	0.531***	0.487***	−0.385***	−0.448***	−0.289***
	(0.013)	(0.014)	(0.013)	(0.022)	(0.023)	(0.016)	(0.018)	(0.024)	(0.025)	(0.016)	(0.018)	(0.014)
Male # Hispanic	0.086***	0.092***	0.116***	0.048***	0.021	0.108***	0.115***	0.047**	0.011	0.105***	0.108***	0.069***
	(0.015)	(0.014)	(0.013)	(0.013)	(0.014)	(0.016)	(0.016)	(0.015)	(0.016)	(0.016)	(0.017)	(0.015)
Asian	−0.339***	−0.463***	−0.480***	−0.420***	−0.040	−0.444***	0.027	−0.310***	0.097*	−0.431***	0.041+	−0.311***
	(0.024)	(0.025)	(0.024)	(0.040)	(0.037)	(0.026)	(0.023)	(0.047)	(0.041)	(0.026)	(0.024)	(0.025)
Male # Asian	−0.046	−0.055+	−0.106***	−0.042	−0.066*	−0.057	−0.091**	−0.054	−0.092**	−0.055	−0.099***	−0.064*
	(0.032)	(0.032)	(0.028)	(0.035)	(0.030)	(0.036)	(0.028)	(0.041)	(0.034)	(0.036)	(0.028)	(0.033)
Native	−0.333***	−0.400***	−0.284***	0.317***	0.235***	−0.575***	−0.647***	0.344***	0.274***	−0.565***	−0.645***	−0.390***
	(0.041)	(0.046)	(0.039)	(0.064)	(0.065)	(0.062)	(0.068)	(0.069)	(0.073)	(0.061)	(0.069)	(0.048)
Male # Native	0.010	0.092+	−0.010	0.039	0.017	0.068	0.107+	−0.024	−0.048	0.059	0.065	−0.001
	(0.053)	(0.050)	(0.044)	(0.040)	(0.042)	(0.059)	(0.059)	(0.049)	(0.053)	(0.061)	(0.059)	(0.054)
Hawaiian	−0.300**	−0.545***	−0.449***	0.223*	0.259**	−0.529***	−0.478***	0.151	0.129	−0.514***	−0.502***	−0.470***
	(0.094)	(0.104)	(0.083)	(0.089)	(0.082)	(0.108)	(0.111)	(0.104)	(0.108)	(0.108)	(0.113)	(0.100)

Table C.14—Continued

Characteristic	Model											
	1a	1b	1c	2a	2b	2c	2d	2i	2j	2k	2l	3a
Male # Hawaiian	0.065	0.174	−0.028	−0.177	−0.181+	0.117	0.252+	−0.177	−0.002	0.144	0.227	0.184
	(0.127)	(0.133)	(0.111)	(0.108)	(0.106)	(0.150)	(0.138)	(0.149)	(0.136)	(0.153)	(0.140)	(0.135)
Multi	−0.080***	−0.154***	−0.059***	0.225***	0.182***	−0.169***	−0.259***	0.273***	0.239***	−0.171***	−0.249***	−0.130***
	(0.021)	(0.021)	(0.018)	(0.024)	(0.025)	(0.023)	(0.024)	(0.029)	(0.030)	(0.023)	(0.025)	(0.022)
Male # multi	0.003	0.054+	0.005	0.016	0.005	0.027	0.054+	0.009	0.002	0.025	0.055+	−0.017
	(0.029)	(0.029)	(0.025)	(0.027)	(0.027)	(0.031)	(0.031)	(0.032)	(0.032)	(0.032)	(0.032)	(0.031)
Other	−0.009	−0.061**	−0.073***	0.180***	0.146***	−0.073**	−0.107***	0.254***	0.205***	−0.058*	−0.127***	−0.079***
	(0.021)	(0.019)	(0.019)	(0.040)	(0.042)	(0.023)	(0.023)	(0.044)	(0.046)	(0.023)	(0.023)	(0.021)
Male # other	0.011	−0.001	0.019	0.026	0.023	−0.002	0.022	0.013	0.020	−0.011	0.026	0.001
	(0.020)	(0.019)	(0.017)	(0.020)	(0.021)	(0.021)	(0.020)	(0.023)	(0.024)	(0.021)	(0.021)	(0.021)
N	**2,280,166**	**2,280,166**	**2,280,166**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,219,066**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.15
Student-Level Regressions: False Negative (Mathematics)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male	−0.068***	−0.068***	−0.002	0.021***	−0.130***	0.085***	−0.097***	0.022***	−0.093***	0.085***	−0.097***	0.027***
	(0.008)	(0.007)	(0.008)	(0.006)	(0.006)	(0.007)	(0.007)	(0.005)	(0.005)	(0.007)	(0.007)	(0.007)
Black	−0.277***	−0.320***	−0.293***	−0.932***	−0.798***	−0.264***	−0.079***	−1.018***	−0.912***	−0.287***	−0.099***	−0.326***
	(0.019)	(0.016)	(0.019)	(0.028)	(0.026)	(0.016)	(0.015)	(0.026)	(0.025)	(0.016)	(0.015)	(0.016)
Male # black	0.036*	0.040*	0.015	−0.003	0.035*	−0.046**	0.030*	−0.019	0.006	−0.048**	0.029*	0.009
	(0.018)	(0.016)	(0.017)	(0.016)	(0.014)	(0.015)	(0.014)	(0.014)	(0.013)	(0.015)	(0.014)	(0.016)
Hispanic	−0.121***	−0.156***	−0.036*	−0.691***	−0.658***	−0.097***	−0.060***	−0.797***	−0.773***	−0.116***	−0.070***	−0.191***
	(0.016)	(0.014)	(0.016)	(0.023)	(0.023)	(0.014)	(0.014)	(0.023)	(0.022)	(0.014)	(0.014)	(0.014)
Male # Hispanic	0.059***	0.062***	0.024+	0.050***	0.081***	0.025+	0.081***	0.052***	0.069***	0.024+	0.078***	0.070***
	(0.015)	(0.014)	(0.015)	(0.012)	(0.012)	(0.013)	(0.013)	(0.011)	(0.011)	(0.013)	(0.013)	(0.014)
Asian	−0.190***	−0.060**	−0.016	0.359***	0.059+	−0.059*	−0.553***	0.311***	0.056+	−0.064**	−0.566***	−0.254***
	(0.027)	(0.023)	(0.025)	(0.029)	(0.033)	(0.024)	(0.028)	(0.027)	(0.030)	(0.023)	(0.028)	(0.025)
Male # Asian	−0.020	0.000	0.034	−0.134***	−0.126***	−0.017	0.027	−0.128***	−0.145***	−0.024	0.023	−0.007
	(0.031)	(0.026)	(0.027)	(0.018)	(0.020)	(0.025)	(0.030)	(0.016)	(0.017)	(0.025)	(0.030)	(0.029)
Native	−0.256***	−0.325***	−0.299***	−0.662***	−0.622***	−0.230***	−0.164***	−0.753***	−0.716***	−0.262***	−0.176***	−0.241***
	(0.059)	(0.048)	(0.059)	(0.090)	(0.083)	(0.045)	(0.043)	(0.088)	(0.085)	(0.045)	(0.042)	(0.049)
Male # Native	−0.046	0.005	0.031	0.052	0.082*	−0.047	0.051	0.065+	0.082*	−0.046	0.042	−0.001
	(0.055)	(0.051)	(0.055)	(0.039)	(0.040)	(0.046)	(0.045)	(0.038)	(0.036)	(0.045)	(0.046)	(0.053)
Hawaiian	−0.398***	−0.360***	−0.229*	−0.642***	−0.655***	−0.282***	−0.256**	−0.672***	−0.668***	−0.303***	−0.256**	−0.224*
	(0.106)	(0.085)	(0.089)	(0.093)	(0.088)	(0.080)	(0.079)	(0.082)	(0.080)	(0.079)	(0.079)	(0.095)

Table C.15—Continued

Characteristic	Model											
	1a	1b	1c	2a	2b	2c	2d	2i	2j	2k	2l	3a
Male # Hawaiian	0.196	0.086	0.106	0.164+	0.148	0.077	0.178	0.072	0.066	0.070	0.141	−0.008
	(0.146)	(0.127)	(0.119)	(0.097)	(0.099)	(0.115)	(0.113)	(0.089)	(0.089)	(0.113)	(0.112)	(0.126)
Multi	−0.152***	−0.121***	−0.183***	−0.263***	−0.214***	−0.152***	−0.059**	−0.313***	−0.276***	−0.150***	−0.071***	−0.147***
	(0.023)	(0.020)	(0.023)	(0.022)	(0.021)	(0.020)	(0.020)	(0.020)	(0.019)	(0.020)	(0.020)	(0.022)
Male # multi	0.006	0.009	0.083**	0.021	0.044*	0.026	0.034	0.023	0.035+	0.021	0.028	0.029
	(0.032)	(0.028)	(0.030)	(0.022)	(0.022)	(0.027)	(0.026)	(0.020)	(0.020)	(0.026)	(0.027)	(0.029)
Other	−0.115***	−0.077***	−0.042+	−0.290***	−0.274***	−0.049*	−0.012	−0.343***	−0.330***	−0.060**	−0.001	−0.043*
	(0.022)	(0.020)	(0.024)	(0.039)	(0.037)	(0.020)	(0.020)	(0.036)	(0.036)	(0.020)	(0.020)	(0.021)
Male # other	0.035	0.029	0.014	0.018	0.006	−0.006	0.006	0.013	0.004	−0.006	0.010	0.035+
	(0.022)	(0.021)	(0.021)	(0.017)	(0.016)	(0.019)	(0.019)	(0.015)	(0.015)	(0.019)	(0.019)	(0.020)
N	**2,280,166**	**2,280,166**	**2,280,166**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,279,643**	**2,219,066**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.16
School-Level Regressions: Bias (Mathematics)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Rural	0.009+	0.012*	0.020**	−0.102***	−0.102***	−0.012+	0.003	−0.120***	−0.118***	−0.010	0.000	−0.018***
	(0.005)	(0.006)	(0.007)	(0.013)	(0.013)	(0.007)	(0.007)	(0.013)	(0.013)	(0.007)	(0.007)	(0.004)
Suburban	0.004	−0.002	0.011+	−0.139***	−0.139***	−0.007	0.005	−0.143***	−0.142***	−0.007	0.005	−0.012***
	(0.005)	(0.005)	(0.006)	(0.012)	(0.012)	(0.005)	(0.006)	(0.012)	(0.012)	(0.005)	(0.006)	(0.003)
Town	0.005	0.016*	0.032***	−0.054***	−0.055***	−0.001	0.007	−0.067***	−0.066***	0.001	0.005	−0.017***
	(0.006)	(0.007)	(0.008)	(0.016)	(0.016)	(0.008)	(0.010)	(0.016)	(0.016)	(0.008)	(0.010)	(0.005)
Constant	0.028***	−0.014***	0.059***	0.178***	0.179***	0.012**	0.001	0.074***	0.075***	0.001	−0.006	0.008**
	(0.003)	(0.004)	(0.004)	(0.010)	(0.010)	(0.004)	(0.005)	(0.009)	(0.009)	(0.004)	(0.005)	(0.003)
***N* schools**	**13,713**	**13,713**	**13,713**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**11,099**
Percentage of students eligible for FRPL	−0.002	0.009	−0.017+	0.949***	0.947***	0.060***	−0.006	0.974***	0.973***	0.063***	−0.002	0.030***
	(0.007)	(0.008)	(0.009)	(0.016)	(0.016)	(0.008)	(0.009)	(0.016)	(0.016)	(0.008)	(0.009)	(0.005)
Constant	0.033***	−0.015***	0.080***	−0.391***	−0.390***	−0.024***	0.007	−0.515***	−0.513***	−0.037***	−0.003	−0.017***
	(0.004)	(0.004)	(0.005)	(0.010)	(0.010)	(0.005)	(0.006)	(0.010)	(0.010)	(0.005)	(0.005)	(0.003)
***N* schools**	**13,713**	**13,713**	**13,713**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**11,099**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.17
School-Level Regressions: Mean Square Error (Mathematics)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Rural	−0.022***	−0.048***	−0.038***	−0.164***	−0.162***	0.050	−0.030	−0.136***	−0.134***	0.045	−0.037	−0.001
	(0.006)	(0.007)	(0.008)	(0.018)	(0.018)	(0.066)	(0.032)	(0.016)	(0.016)	(0.066)	(0.032)	(0.003)
Suburban	−0.010	−0.037***	−0.020*	−0.110***	−0.108***	−0.024**	−0.046*	−0.063***	−0.063***	−0.025***	−0.047*	−0.010***
	(0.007)	(0.007)	(0.009)	(0.015)	(0.016)	(0.008)	(0.023)	(0.014)	(0.014)	(0.008)	(0.023)	(0.003)
Town	−0.022***	−0.029***	−0.021*	−0.142***	−0.140***	−0.003	0.025	−0.116***	−0.113***	−0.005	0.020	0.007
	(0.006)	(0.009)	(0.010)	(0.022)	(0.022)	(0.016)	(0.045)	(0.020)	(0.020)	(0.015)	(0.045)	(0.005)
Constant	0.216***	0.321***	0.392***	0.634***	0.633***	0.299***	0.335***	0.595***	0.593***	0.298***	0.334***	0.230***
	(0.004)	(0.005)	(0.006)	(0.012)	(0.012)	(0.005)	(0.022)	(0.011)	(0.011)	(0.005)	(0.022)	(0.002)
***N* schools**	**13,713**	**13,713**	**13,713**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**11,099**
Percentage of students eligible for FRPL	0.048***	0.118***	0.118***	0.167***	0.164***	0.090***	0.103**	−0.046*	−0.043*	0.095***	0.108**	0.062***
	(0.008)	(0.010)	(0.012)	(0.022)	(0.023)	(0.016)	(0.037)	(0.021)	(0.021)	(0.018)	(0.038)	(0.005)
Constant	0.180***	0.231***	0.312***	0.452***	0.453***	0.257***	0.261***	0.548***	0.547***	0.251***	0.255***	0.194***
	(0.006)	(0.006)	(0.007)	(0.014)	(0.014)	(0.014)	(0.018)	(0.013)	(0.013)	(0.014)	(0.018)	(0.003)
***N* schools**	**13,713**	**13,713**	**13,713**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**11,099**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.18
School-Level Regressions: Exact Agreement (Mathematics)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Rural	−0.018***	−0.021***	−0.025***	0.008+	0.007+	−0.022***	−0.024***	−0.005	−0.006	−0.021***	−0.023***	−0.032***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)
Suburban	−0.013***	−0.017***	−0.019***	−0.013***	−0.014***	−0.016***	−0.019***	−0.028***	−0.027***	−0.017***	−0.018***	−0.020***
	(0.002)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)
Town	−0.017***	−0.023***	−0.022***	0.018***	0.019***	−0.023***	−0.029***	0.006	0.005	−0.025***	−0.026***	−0.026***
	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.002)
Constant	0.881***	0.862***	0.841***	0.767***	0.768***	0.861***	0.860***	0.775***	0.776***	0.861***	0.860***	0.873***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.001)
***N* schools**	**13,713**	**13,713**	**13,713**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**11,099**
Percentage of students eligible for FRPL	0.047***	0.050***	0.053***	0.113***	0.112***	0.052***	0.051***	0.188***	0.184***	0.056***	0.052***	0.058***
	(0.004)	(0.004)	(0.004)	(0.006)	(0.006)	(0.004)	(0.004)	(0.006)	(0.006)	(0.004)	(0.004)	(0.003)
Constant	0.845***	0.822***	0.798***	0.709***	0.709***	0.820***	0.818***	0.667***	0.671***	0.817***	0.818***	0.826***
	(0.002)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)
***N* schools**	**13,713**	**13,713**	**13,713**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**11,099**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.19
School-Level Regressions: False Positive (Mathematics)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Rural	0.014***	0.012***	0.018***	−0.031***	−0.031***	0.008***	0.014***	−0.029***	−0.029***	0.007**	0.011***	0.013***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Suburban	0.010***	0.008***	0.014***	−0.023***	−0.022***	0.009***	0.016***	−0.019***	−0.019***	0.008***	0.014***	0.009***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Town	0.010***	0.013***	0.018***	−0.022***	−0.022***	0.009**	0.014***	−0.020***	−0.018***	0.009**	0.012***	0.010***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)
Constant	0.061***	0.065***	0.093***	0.108***	0.108***	0.055***	0.053***	0.079***	0.079***	0.054***	0.053***	0.060***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
***N* schools**	**13,713**	**13,713**	**13,713**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**11,099**
Percentage of students eligible for FRPL	−0.029***	−0.026***	−0.040***	0.146***	0.145***	−0.046***	−0.063***	0.119***	0.120***	−0.044***	−0.060***	−0.039***
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
Constant	0.084***	0.086***	0.124***	0.015***	0.016***	0.085***	0.096***	0.001	0.001	0.082***	0.093***	0.087***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)
***N* schools**	**13,713**	**13,713**	**13,713**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**11,099**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.20
School-Level Regressions: False Negative (Mathematics)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Rural	0.004*	0.009***	0.007**	0.024***	0.024***	0.014***	0.011***	0.034***	0.035***	0.015***	0.012***	0.019***
	(0.002)	(0.002)	(0.002)	(0.005)	(0.005)	(0.003)	(0.003)	(0.005)	(0.005)	(0.003)	(0.003)	(0.002)
Suburban	0.004*	0.010***	0.005*	0.036***	0.036***	0.007**	0.003	0.047***	0.047***	0.009***	0.004+	0.010***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.002)	(0.002)	(0.004)	(0.004)	(0.002)	(0.002)	(0.001)
Town	0.007*	0.010**	0.004	0.004	0.003	0.014***	0.015***	0.013*	0.013*	0.016***	0.014***	0.016***
	(0.003)	(0.003)	(0.003)	(0.005)	(0.005)	(0.003)	(0.004)	(0.006)	(0.006)	(0.003)	(0.003)	(0.002)
Constant	0.058***	0.072***	0.067***	0.124***	0.124***	0.084***	0.087***	0.147***	0.145***	0.086***	0.087***	0.067***
	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.001)
***N* schools**	**13,713**	**13,713**	**13,713**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**11,099**
Percentage of students eligible for FRPL	−0.018***	−0.024***	−0.014***	−0.259***	−0.257***	−0.006+	0.012***	−0.307***	−0.304***	−0.012***	0.008*	−0.020***
	(0.003)	(0.003)	(0.003)	(0.006)	(0.006)	(0.003)	(0.003)	(0.006)	(0.006)	(0.003)	(0.003)	(0.002)
Constant	0.071***	0.092***	0.078***	0.277***	0.275***	0.095***	0.086***	0.331***	0.328***	0.101***	0.089***	0.087***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.002)	(0.002)	(0.004)	(0.004)	(0.002)	(0.002)	(0.001)
***N* schools**	**13,713**	**13,713**	**13,713**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**13,702**	**11,099**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.21
Student-Level Regressions: Bias (Reading)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male	−0.002*	0.015***	−0.002	0.092***	0.016***	0.090***	0.011***	0.094***	0.017***	0.090***	0.012***	0.055***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Black	−0.011***	−0.012***	0.003	0.306***	0.275***	0.029***	−0.012***	0.316***	0.288***	0.032***	−0.011***	−0.002
	(0.002)	(0.002)	(0.003)	(0.007)	(0.008)	(0.002)	(0.002)	(0.008)	(0.008)	(0.002)	(0.002)	(0.002)
Male # black	−0.002	0.003	0.006*	0.012***	0.014***	0.007**	0.008***	0.013***	0.014***	0.006**	0.007**	−0.004+
	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
Hispanic	−0.013***	−0.053***	−0.064***	0.237***	0.216***	0.021***	−0.009***	0.267***	0.248***	0.025***	−0.006*	0.003
	(0.002)	(0.002)	(0.003)	(0.007)	(0.007)	(0.002)	(0.002)	(0.007)	(0.007)	(0.002)	(0.002)	(0.002)
Male # Hispanic	0.002	0.001	0.015***	0.004+	0.002	0.004+	0.001	0.005*	0.003	0.004+	0.001	−0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Asian	0.000	−0.052***	−0.036***	−0.146***	−0.111***	−0.038***	0.004	−0.131***	−0.098***	−0.037***	0.003	0.012***
	(0.002)	(0.003)	(0.005)	(0.010)	(0.011)	(0.003)	(0.003)	(0.011)	(0.011)	(0.003)	(0.003)	(0.003)
Male # Asian	−0.004	−0.016***	−0.011**	−0.005	−0.006	−0.012***	−0.012***	−0.006	−0.006	−0.012***	−0.012***	−0.014***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)
Native	0.013*	−0.019**	0.007	0.310***	0.284***	0.035***	−0.013+	0.290***	0.264***	0.037***	−0.013+	0.006
	(0.005)	(0.007)	(0.008)	(0.027)	(0.026)	(0.007)	(0.007)	(0.026)	(0.026)	(0.007)	(0.007)	(0.006)
Male # Native	0.007	0.015*	0.017*	0.012	0.015	0.019**	0.016*	0.012	0.014	0.019**	0.018**	0.012+
	(0.007)	(0.007)	(0.008)	(0.010)	(0.010)	(0.007)	(0.007)	(0.009)	(0.009)	(0.007)	(0.007)	(0.006)
Hawaiian	0.013	−0.041***	−0.019	0.199***	0.179***	0.025*	−0.004	0.190***	0.171***	0.030*	−0.002	0.005
	(0.011)	(0.012)	(0.015)	(0.026)	(0.025)	(0.012)	(0.012)	(0.027)	(0.027)	(0.012)	(0.012)	(0.011)

Table C.21—Continued

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male # Hawaiian	−0.002	−0.003	0.007	0.001	0.005	−0.004	−0.006	0.004	0.009	−0.004	−0.007	0.016
	(0.014)	(0.016)	(0.018)	(0.019)	(0.019)	(0.015)	(0.016)	(0.019)	(0.019)	(0.015)	(0.016)	(0.015)
Multi	0.008**	−0.002	0.012***	0.072***	0.071***	0.003	−0.002	0.075***	0.074***	0.002	−0.001	−0.006*
	(0.003)	(0.003)	(0.003)	(0.006)	(0.006)	(0.003)	(0.003)	(0.006)	(0.006)	(0.003)	(0.003)	(0.003)
Male # multi	0.003	0.000	0.005	−0.002	−0.001	0.006	0.006	0.000	0.000	0.006	0.006	0.004
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)
Other	0.002	−0.011**	−0.015**	0.092***	0.083***	0.007+	−0.003	0.110***	0.100***	0.011**	−0.003	0.004
	(0.003)	(0.004)	(0.005)	(0.012)	(0.012)	(0.004)	(0.004)	(0.012)	(0.012)	(0.004)	(0.003)	(0.003)
Male # other	−0.004+	−0.007*	−0.005	−0.005	−0.002	−0.005+	−0.003	−0.005	−0.003	−0.006+	−0.003	−0.008**
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
N	**2,280,084**	**2,280,084**	**2,280,084**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,433**	**2,279,422**	**2,279,433**	**2,279,422**	**2,218,961**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.22
Student-Level Regressions: Mean Square Error (Reading)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male	0.049***	0.072***	0.080***	0.046***	0.060***	0.059***	0.061***	0.025***	0.056***	0.058***	0.061***	0.047***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Black	0.050***	0.074***	0.061***	0.112***	0.131***	0.056***	0.062***	0.044***	0.070***	0.055***	0.062***	0.038***
	(0.002)	(0.003)	(0.003)	(0.006)	(0.006)	(0.002)	(0.003)	(0.005)	(0.005)	(0.002)	(0.003)	(0.002)
Male # black	0.023***	0.028***	0.025***	0.089***	0.042***	0.029***	0.024***	0.090***	0.040***	0.029***	0.024***	0.016***
	(0.002)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.002)
Hispanic	0.027***	0.044***	0.032***	0.041***	0.060***	0.032***	0.033***	−0.008*	0.016***	0.031***	0.033***	0.019***
	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)	(0.002)	(0.002)	(0.004)	(0.004)	(0.002)	(0.002)	(0.002)
Male # Hispanic	0.002	0.003	−0.002	0.052***	0.016***	0.009***	0.004*	0.058***	0.016***	0.010***	0.004+	0.003+
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
Asian	−0.035***	−0.040***	−0.044***	0.059***	0.036***	−0.034***	−0.036***	0.081***	0.054***	−0.033***	−0.037***	−0.040***
	(0.002)	(0.003)	(0.003)	(0.007)	(0.007)	(0.002)	(0.003)	(0.009)	(0.008)	(0.002)	(0.003)	(0.002)
Male # Asian	−0.024***	−0.040***	−0.040***	−0.048***	−0.026***	−0.039***	−0.033***	−0.046***	−0.025***	−0.039***	−0.032***	−0.017***
	(0.002)	(0.003)	(0.004)	(0.005)	(0.004)	(0.003)	(0.003)	(0.005)	(0.004)	(0.003)	(0.003)	(0.003)
Native	0.036***	0.053***	0.047***	0.192***	0.216***	0.061**	0.065***	0.109***	0.139***	0.061**	0.065***	0.036***
	(0.005)	(0.007)	(0.007)	(0.023)	(0.025)	(0.019)	(0.009)	(0.018)	(0.019)	(0.019)	(0.009)	(0.006)
Male # Native	0.016*	0.013	0.012	0.083***	0.033*	0.005	0.012	0.077***	0.031*	0.009	0.011	0.012+
	(0.007)	(0.009)	(0.009)	(0.015)	(0.014)	(0.017)	(0.009)	(0.014)	(0.012)	(0.014)	(0.009)	(0.007)
Hawaiian	0.015	0.028*	0.007	0.038*	0.064***	0.009	0.031**	0.000	0.032*	0.008	0.033**	0.015
	(0.009)	(0.013)	(0.015)	(0.016)	(0.016)	(0.011)	(0.011)	(0.015)	(0.015)	(0.010)	(0.012)	(0.010)

Table C.22—Continued

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male # Hawaiian	−0.007	−0.025	−0.008	0.031	0.004	−0.011	−0.008	0.033+	0.011	−0.010	−0.009	0.001
	(0.014)	(0.019)	(0.020)	(0.020)	(0.020)	(0.014)	(0.015)	(0.019)	(0.019)	(0.014)	(0.016)	(0.014)
Multi	0.018***	0.025***	0.021***	0.050***	0.063***	0.020***	0.022***	0.032***	0.045***	0.020***	0.022***	0.012***
	(0.003)	(0.003)	(0.004)	(0.005)	(0.005)	(0.003)	(0.003)	(0.005)	(0.005)	(0.003)	(0.003)	(0.002)
Male # multi	0.006	0.006	0.005	0.022***	0.012*	0.004	0.005	0.024***	0.013*	0.004	0.004	0.003
	(0.004)	(0.005)	(0.006)	(0.006)	(0.005)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)
Other	0.011***	0.018***	0.010**	0.008	0.016*	0.019***	0.020***	−0.019**	−0.008	0.018***	0.020***	0.010***
	(0.002)	(0.003)	(0.004)	(0.007)	(0.007)	(0.005)	(0.003)	(0.007)	(0.007)	(0.005)	(0.003)	(0.002)
Male # other	0.001	0.000	0.002	0.019***	0.006+	0.001	−0.001	0.024***	0.007*	0.002	−0.001	0.003
	(0.003)	(0.003)	(0.004)	(0.005)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)
N	**2,280,084**	**2,280,084**	**2,280,084**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,433**	**2,279,422**	**2,279,433**	**2,279,422**	**2,218,961**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.23
Student-Level Regressions: Exact Agreement (Reading)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male	0.010+	0.007	−0.006	0.104***	0.046***	0.026***	0.008	0.156***	0.064***	0.031***	0.006	0.013*
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Black	0.172***	0.182***	0.157***	0.176***	0.115***	0.206***	0.189***	0.380***	0.309***	0.210***	0.185***	0.238***
	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)	(0.011)	(0.011)	(0.010)
Male # black	0.171***	0.145***	0.174***	0.065***	0.195***	0.169***	0.176***	0.048***	0.180***	0.166***	0.182***	0.170***
	(0.012)	(0.011)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)	(0.010)	(0.010)	(0.011)	(0.011)	(0.011)
Hispanic	0.151***	0.180***	0.183***	0.247***	0.185***	0.195***	0.177***	0.411***	0.341***	0.198***	0.179***	0.209***
	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)	(0.010)	(0.010)	(0.010)
Male # Hispanic	0.129***	0.115***	0.126***	0.017+	0.119***	0.122***	0.126***	−0.011	0.101***	0.120***	0.125***	0.109***
	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)
Asian	0.177***	0.183***	0.229***	−0.125***	−0.049**	0.168***	0.182***	−0.168***	−0.098***	0.166***	0.182***	0.207***
	(0.017)	(0.016)	(0.016)	(0.020)	(0.019)	(0.016)	(0.016)	(0.021)	(0.020)	(0.016)	(0.016)	(0.017)
Male # Asian	−0.025	−0.040*	−0.038*	−0.001	−0.066***	−0.025	−0.031	−0.023	−0.069***	−0.025	−0.029	−0.035+
	(0.020)	(0.019)	(0.019)	(0.017)	(0.017)	(0.020)	(0.020)	(0.016)	(0.016)	(0.020)	(0.020)	(0.020)
Native	0.248***	0.337***	0.278***	0.239***	0.181***	0.345***	0.286***	0.412***	0.343***	0.333***	0.294***	0.309***
	(0.031)	(0.032)	(0.032)	(0.034)	(0.033)	(0.034)	(0.033)	(0.042)	(0.040)	(0.034)	(0.033)	(0.035)
Male # Native	0.195***	0.139***	0.143***	0.102***	0.192***	0.158***	0.179***	0.073*	0.170***	0.166***	0.166***	0.177***
	(0.036)	(0.036)	(0.035)	(0.030)	(0.032)	(0.038)	(0.037)	(0.031)	(0.033)	(0.038)	(0.038)	(0.039)
Hawaiian	0.368***	0.292***	0.318***	0.333***	0.211***	0.349***	0.303***	0.489***	0.418***	0.350***	0.306***	0.340***
	(0.067)	(0.067)	(0.068)	(0.061)	(0.061)	(0.070)	(0.071)	(0.062)	(0.062)	(0.069)	(0.072)	(0.071)

Table C.23—Continued

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male # Hawaiian	0.023	0.203*	0.143+	0.085	0.242**	0.171*	0.126	0.087	0.180*	0.231**	0.133	0.179*
	(0.093)	(0.091)	(0.085)	(0.079)	(0.080)	(0.087)	(0.086)	(0.082)	(0.082)	(0.088)	(0.086)	(0.089)
Multi	0.071***	0.070***	0.060***	0.038**	0.000	0.079***	0.083***	0.107***	0.075***	0.082***	0.083***	0.091***
	(0.015)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Male # multi	0.042*	0.041*	0.044*	0.001	0.037*	0.058**	0.026	–0.022	0.011	0.051*	0.028	0.050*
	(0.021)	(0.019)	(0.019)	(0.018)	(0.017)	(0.020)	(0.020)	(0.018)	(0.018)	(0.020)	(0.020)	(0.020)
Other	0.022+	0.023*	0.026*	0.092***	0.078***	0.029*	0.011	0.169***	0.145***	0.031**	0.012	0.024*
	(0.011)	(0.011)	(0.011)	(0.017)	(0.016)	(0.012)	(0.012)	(0.020)	(0.019)	(0.012)	(0.012)	(0.012)
Male # other	0.052***	0.043**	0.067***	0.007	0.034**	0.041**	0.048***	–0.011	0.027*	0.042**	0.049***	0.056***
	(0.014)	(0.013)	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)	(0.013)	(0.013)	(0.013)	(0.014)	(0.015)
N	**2,280,084**	**2,280,084**	**2,280,084**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,218,961**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.24
Student-Level Regressions: False Positive (Reading)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male	−0.028***	0.029***	0.003	0.159***	−0.144***	0.220***	−0.095***	0.173***	−0.146***	0.222***	−0.091***	0.095***
	(0.007)	(0.006)	(0.006)	(0.008)	(0.007)	(0.007)	(0.007)	(0.009)	(0.009)	(0.007)	(0.007)	(0.007)
Black	−0.200***	−0.205***	−0.104***	0.598***	0.471***	−0.443***	−0.603***	0.678***	0.551***	−0.433***	−0.589***	−0.363***
	(0.012)	(0.013)	(0.012)	(0.021)	(0.021)	(0.018)	(0.018)	(0.023)	(0.023)	(0.019)	(0.019)	(0.014)
Male # black	−0.154***	−0.122***	−0.170***	−0.224***	−0.187***	−0.188***	−0.221***	−0.234***	−0.205***	−0.190***	−0.230***	−0.199***
	(0.015)	(0.015)	(0.013)	(0.013)	(0.014)	(0.018)	(0.019)	(0.016)	(0.016)	(0.018)	(0.019)	(0.016)
Hispanic	−0.184***	−0.297***	−0.275***	0.392***	0.302***	−0.458***	−0.552***	0.521***	0.438***	−0.446***	−0.543***	−0.308***
	(0.011)	(0.012)	(0.012)	(0.021)	(0.021)	(0.016)	(0.016)	(0.023)	(0.023)	(0.016)	(0.015)	(0.013)
Male # Hispanic	−0.100***	−0.089***	−0.086***	−0.116***	−0.096***	−0.081***	−0.142***	−0.099***	−0.090***	−0.081***	−0.137***	−0.094***
	(0.013)	(0.013)	(0.012)	(0.013)	(0.013)	(0.016)	(0.016)	(0.015)	(0.015)	(0.016)	(0.016)	(0.014)
Asian	−0.176***	−0.343***	−0.312***	−0.252***	−0.136***	−0.136***	−0.015	−0.144***	−0.040	−0.132***	−0.016	−0.067**
	(0.021)	(0.021)	(0.021)	(0.036)	(0.034)	(0.023)	(0.020)	(0.043)	(0.038)	(0.023)	(0.020)	(0.020)
Male # Asian	0.013	0.019	0.007	0.002	0.036	−0.015	0.018	−0.016	0.024	−0.016	0.012	−0.017
	(0.027)	(0.027)	(0.025)	(0.030)	(0.029)	(0.028)	(0.026)	(0.036)	(0.034)	(0.028)	(0.026)	(0.026)
Native	−0.234***	−0.363***	−0.254***	0.356***	0.253***	−0.515***	−0.597***	0.382***	0.293***	−0.499***	−0.606***	−0.358***
	(0.037)	(0.044)	(0.038)	(0.057)	(0.058)	(0.060)	(0.062)	(0.063)	(0.063)	(0.060)	(0.065)	(0.052)
Male # Native	−0.155**	−0.099*	−0.124**	−0.220***	−0.143***	−0.115*	−0.065	−0.193***	−0.162***	−0.119*	−0.05	−0.148**
	(0.048)	(0.046)	(0.042)	(0.041)	(0.042)	(0.058)	(0.060)	(0.047)	(0.047)	(0.059)	(0.060)	(0.054)
Hawaiian	−0.311***	−0.260***	−0.260**	0.212*	0.223**	−0.460***	−0.416***	0.336**	0.262**	−0.453***	−0.434***	−0.384***
	(0.079)	(0.078)	(0.081)	(0.096)	(0.085)	(0.099)	(0.095)	(0.109)	(0.101)	(0.098)	(0.099)	(0.090)

Table C.24—Continued

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male # Hawaiian	−0.094	−0.231*	−0.196+	−0.244*	−0.226*	−0.113	−0.023	−0.441**	−0.214	−0.231+	−0.041	−0.127
	(0.123)	(0.117)	(0.109)	(0.115)	(0.110)	(0.135)	(0.129)	(0.144)	(0.137)	(0.140)	(0.130)	(0.130)
Multi	−0.037*	−0.093***	−0.031+	0.202***	0.191***	−0.210***	−0.222***	0.246***	0.246***	−0.209***	−0.222***	−0.145***
	(0.019)	(0.019)	(0.017)	(0.024)	(0.023)	(0.023)	(0.022)	(0.028)	(0.026)	(0.023)	(0.022)	(0.020)
Male # multi	−0.049+	−0.017	−0.029	−0.068**	−0.036	−0.006	0.022	−0.042	−0.013	−0.003	0.023	−0.040
	(0.027)	(0.026)	(0.023)	(0.026)	(0.026)	(0.030)	(0.029)	(0.031)	(0.031)	(0.030)	(0.029)	(0.028)
Other	−0.025	−0.062***	−0.055**	0.181***	0.126***	−0.126***	−0.130***	0.271***	0.215***	−0.112***	−0.134***	−0.064***
	(0.015)	(0.017)	(0.017)	(0.038)	(0.036)	(0.021)	(0.020)	(0.041)	(0.039)	(0.022)	(0.021)	(0.018)
Male # other	−0.029+	−0.038*	−0.054**	−0.046*	−0.013	−0.019	−0.038+	−0.061*	−0.043+	−0.027	−0.039+	−0.049*
	(0.018)	(0.017)	(0.017)	(0.020)	(0.020)	(0.021)	(0.020)	(0.024)	(0.023)	(0.021)	(0.020)	(0.019)
N	**2,280,084**	**2,280,084**	**2,280,084**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,218,961**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.25
Student-Level Regressions: False Negative (Reading)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male	0.013+	−0.046***	0.008	−0.215***	0.013*	−0.214***	0.068***	−0.228***	−0.032***	−0.216***	0.067***	−0.120***
	(0.007)	(0.007)	(0.007)	(0.005)	(0.005)	(0.006)	(0.007)	(0.005)	(0.005)	(0.006)	(0.006)	(0.007)
Black	−0.109***	−0.121***	−0.194***	−0.680***	−0.567***	−0.048***	0.119***	−0.764***	−0.679***	−0.064***	0.111***	−0.094***
	(0.014)	(0.013)	(0.014)	(0.023)	(0.023)	(0.012)	(0.012)	(0.021)	(0.022)	(0.011)	(0.012)	(0.012)
Male # black	−0.165***	−0.142***	−0.141***	−0.039**	−0.098***	−0.079***	−0.175***	−0.049***	−0.094***	−0.079***	−0.177***	−0.098***
	(0.016)	(0.015)	(0.016)	(0.014)	(0.014)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.015)
Hispanic	−0.086***	−0.030*	−0.025+	−0.598***	−0.511***	−0.027*	0.110***	−0.706***	−0.643***	−0.042***	0.098***	−0.087***
	(0.013)	(0.012)	(0.013)	(0.020)	(0.020)	(0.011)	(0.011)	(0.019)	(0.020)	(0.011)	(0.011)	(0.012)
Male # Hispanic	−0.142***	−0.114***	−0.149***	−0.050***	−0.071***	−0.070***	−0.138***	−0.042***	−0.061***	−0.072***	−0.136***	−0.088***
	(0.014)	(0.013)	(0.014)	(0.012)	(0.012)	(0.012)	(0.012)	(0.011)	(0.011)	(0.012)	(0.012)	(0.014)
Asian	−0.144***	0.004	−0.074***	0.241***	0.133***	−0.164***	−0.321***	0.215***	0.124***	−0.164***	−0.316***	−0.315***
	(0.023)	(0.020)	(0.022)	(0.027)	(0.029)	(0.021)	(0.024)	(0.026)	(0.027)	(0.021)	(0.024)	(0.023)
Male # Asian	0.033	0.062*	0.063*	0.039*	0.058**	0.054*	0.059*	0.048**	0.069***	0.053*	0.061*	0.070*
	(0.028)	(0.025)	(0.026)	(0.018)	(0.018)	(0.025)	(0.027)	(0.016)	(0.017)	(0.024)	(0.027)	(0.028)
Native	−0.220***	−0.247***	−0.245***	−0.552***	−0.454***	−0.204***	−0.021	−0.634***	−0.558***	−0.202***	−0.032	−0.215***
	(0.043)	(0.038)	(0.045)	(0.074)	(0.070)	(0.036)	(0.036)	(0.071)	(0.070)	(0.037)	(0.036)	(0.038)
Male # Native	−0.214***	−0.159**	−0.141*	−0.085*	−0.160***	−0.121**	−0.250***	−0.086*	−0.130***	−0.131**	−0.238***	−0.169***
	(0.053)	(0.050)	(0.055)	(0.041)	(0.041)	(0.044)	(0.043)	(0.039)	(0.038)	(0.045)	(0.043)	(0.050)
Hawaiian	−0.375***	−0.270**	−0.326***	−0.588***	−0.477***	−0.239**	−0.162+	−0.714***	−0.644***	−0.248**	−0.157+	−0.245**
	(0.103)	(0.089)	(0.092)	(0.090)	(0.088)	(0.081)	(0.083)	(0.084)	(0.084)	(0.081)	(0.081)	(0.093)

Table C.25—Continued

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Male # Hawaiian	0.062	−0.140	−0.044	−0.025	−0.176+	−0.158	−0.197+	0.025	−0.119	−0.172	−0.194+	−0.196+
	(0.129)	(0.118)	(0.114)	(0.097)	(0.097)	(0.107)	(0.107)	(0.089)	(0.090)	(0.106)	(0.107)	(0.119)
Multi	−0.098***	−0.032+	−0.085***	−0.149***	−0.118***	0.011	0.045*	−0.193***	−0.180***	0.003	0.044*	−0.026
	(0.021)	(0.019)	(0.020)	(0.019)	(0.020)	(0.017)	(0.018)	(0.018)	(0.019)	(0.017)	(0.018)	(0.018)
Male # multi	−0.024	−0.057*	−0.053+	0.008	−0.016	−0.055*	−0.070**	0.013	0.003	−0.045+	−0.072**	−0.040
	(0.029)	(0.027)	(0.028)	(0.021)	(0.021)	(0.025)	(0.025)	(0.020)	(0.020)	(0.024)	(0.025)	(0.026)
Other	−0.013	0.024	0.017	−0.211***	−0.189***	0.035*	0.091***	−0.275***	−0.255***	0.021	0.090***	0.017
	(0.017)	(0.016)	(0.018)	(0.032)	(0.033)	(0.015)	(0.016)	(0.031)	(0.032)	(0.015)	(0.016)	(0.016)
Male # other	−0.068***	−0.036+	−0.067***	−0.021	−0.026+	−0.026	−0.059***	−0.001	−0.006	−0.026	−0.060***	−0.045*
	(0.020)	(0.019)	(0.019)	(0.015)	(0.015)	(0.017)	(0.017)	(0.015)	(0.014)	(0.017)	(0.017)	(0.019)
N	**2,280,084**	**2,280,084**	**2,280,084**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,279,561**	**2,218,961**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.26
School-Level Regressions: Bias (Reading)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Rural	0.003	0.027***	0.013+	−0.083***	−0.082***	0.004	0.015*	−0.100***	−0.099***	0.005	0.014*	0.009*
	(0.005)	(0.006)	(0.007)	(0.012)	(0.012)	(0.007)	(0.007)	(0.012)	(0.012)	(0.007)	(0.007)	(0.004)
Suburban	0.013**	0.012*	0.026***	−0.116***	−0.116***	0.001	0.012*	−0.119***	−0.119***	0.000	0.011*	0.008*
	(0.005)	(0.005)	(0.006)	(0.012)	(0.012)	(0.006)	(0.006)	(0.012)	(0.012)	(0.006)	(0.006)	(0.004)
Town	0.017**	0.026***	0.024**	−0.050***	−0.051***	0.005	0.005	−0.062***	−0.062***	0.005	0.005	0.006
	(0.006)	(0.008)	(0.008)	(0.015)	(0.015)	(0.008)	(0.008)	(0.015)	(0.015)	(0.008)	(0.008)	(0.004)
Constant	0.019***	−0.001	0.058***	0.172***	0.173***	0.017***	0.009*	0.084***	0.084***	0.011*	0.005	0.009***
	(0.003)	(0.004)	(0.005)	(0.009)	(0.009)	(0.005)	(0.004)	(0.009)	(0.009)	(0.005)	(0.004)	(0.003)
***N* schools**	**13,711**	**13,711**	**13,711**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**11,097**
Percentage of students eligible for FRPL	−0.029***	−0.031***	−0.064***	0.873***	0.871***	0.050***	−0.030**	0.896***	0.894***	0.053***	−0.028**	−0.013*
	(0.007)	(0.008)	(0.009)	(0.015)	(0.015)	(0.009)	(0.010)	(0.015)	(0.015)	(0.009)	(0.010)	(0.005)
Constant	0.041***	0.030***	0.106***	−0.345***	−0.343***	−0.007	0.033***	−0.452***	−0.450***	−0.015**	0.028***	0.022***
	(0.004)	(0.005)	(0.005)	(0.009)	(0.009)	(0.005)	(0.006)	(0.009)	(0.009)	(0.006)	(0.006)	(0.003)
***N* schools**	**13,711**	**13,711**	**13,711**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**11,097**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.27
School-Level Regressions: Mean Square Error (Reading)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Rural	−0.019*	−0.046***	−0.024**	−0.136***	−0.134***	−0.002	0.041	−0.121***	−0.120***	0.006	0.042	0.005
	(0.007)	(0.008)	(0.008)	(0.018)	(0.018)	(0.035)	(0.056)	(0.016)	(0.016)	(0.036)	(0.056)	(0.004)
Suburban	−0.016*	−0.051***	−0.022**	−0.115***	−0.111***	−0.056**	−0.034***	−0.084***	−0.081***	−0.052**	−0.034***	−0.011**
	(0.007)	(0.008)	(0.009)	(0.016)	(0.017)	(0.017)	(0.009)	(0.015)	(0.015)	(0.017)	(0.009)	(0.004)
Town	−0.002	−0.020+	0.004	−0.120***	−0.119***	0.003	0.017	−0.100***	−0.099***	0.006	0.015	0.009+
	(0.008)	(0.011)	(0.011)	(0.021)	(0.021)	(0.032)	(0.014)	(0.019)	(0.019)	(0.032)	(0.014)	(0.005)
Constant	0.280***	0.389***	0.450***	0.633***	0.631***	0.359***	0.345***	0.602***	0.600***	0.355***	0.344***	0.274***
	(0.005)	(0.006)	(0.006)	(0.013)	(0.013)	(0.017)	(0.007)	(0.012)	(0.012)	(0.016)	(0.007)	(0.003)
***N* schools**	**13,711**	**13,711**	**13,711**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**11,097**
Percentage of students eligible for FRPL	0.098***	0.162***	0.151***	0.271***	0.262***	0.094**	0.061	0.105***	0.099***	0.085*	0.064	0.100***
	(0.010)	(0.011)	(0.012)	(0.023)	(0.024)	(0.033)	(0.080)	(0.021)	(0.022)	(0.033)	(0.080)	(0.006)
Constant	0.219***	0.274***	0.359***	0.405***	0.410***	0.291***	0.314***	0.476***	0.480***	0.296***	0.311***	0.219***
	(0.006)	(0.006)	(0.007)	(0.013)	(0.015)	(0.025)	(0.055)	(0.012)	(0.013)	(0.025)	(0.055)	(0.003)
***N* schools**	**13,711**	**13,711**	**13,711**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**11,097**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.28
School-Level Regressions: Exact Agreement (Reading)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Rural	−0.018***	−0.018***	−0.025***	0.011**	0.011**	−0.019***	−0.022***	0.002	0.001	−0.018***	−0.020***	−0.033***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)
Suburban	−0.015***	−0.012***	−0.018***	−0.002	−0.002	−0.014***	−0.015***	−0.016***	−0.016***	−0.014***	−0.012***	−0.020***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)
Town	−0.024***	−0.021***	−0.029***	0.015**	0.015**	−0.024***	−0.023***	0.007	0.006	−0.023***	−0.022***	−0.029***
	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)	(0.003)
Constant	0.859***	0.837***	0.826***	0.752***	0.752***	0.842***	0.840***	0.758***	0.759***	0.841***	0.838***	0.852***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
***N* schools**	**13,711**	**13,711**	**13,711**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**11,097**
Percentage of students eligible for FRPL	0.053***	0.057***	0.057***	0.113***	0.112***	0.059***	0.062***	0.182***	0.181***	0.059***	0.058***	0.066***
	(0.004)	(0.004)	(0.004)	(0.006)	(0.006)	(0.004)	(0.005)	(0.006)	(0.006)	(0.004)	(0.004)	(0.003)
Constant	0.819***	0.797***	0.781***	0.698***	0.698***	0.799***	0.794***	0.660***	0.661***	0.799***	0.797***	0.801***
	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)
***N* schools**	**13,711**	**13,711**	**13,711**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**11,097**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.29
School-Level Regressions: False Positive (Reading)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Rural	0.015***	0.016***	0.019***	−0.026***	−0.026***	0.014***	0.020***	−0.025***	−0.025***	0.013***	0.018***	0.020***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Suburban	0.012***	0.007***	0.015***	−0.024***	−0.025***	0.013***	0.017***	−0.020***	−0.019***	0.012***	0.014***	0.015***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Town	0.018***	0.014***	0.021***	−0.019***	−0.019***	0.013***	0.014***	−0.017***	−0.017***	0.013***	0.013***	0.015***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)
Constant	0.073***	0.080***	0.095***	0.109***	0.110***	0.058***	0.057***	0.081***	0.081***	0.056***	0.058***	0.066***
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
***N* schools**	**13,711**	**13,711**	**13,711**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**11,097**
Percentage of students eligible for FRPL	−0.035***	−0.036***	−0.045***	0.143***	0.143***	−0.056***	−0.077***	0.117***	0.117***	−0.054***	−0.073***	−0.056***
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)
Constant	0.101***	0.107***	0.131***	0.018***	0.018***	0.096***	0.110***	0.006***	0.006***	0.093***	0.106***	0.107***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)
***N* schools**	**13,711**	**13,711**	**13,711**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**11,097**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table C.30
School-Level Regressions: False Negative (Reading)

	Model											
Characteristic	**1a**	**1b**	**1c**	**2a**	**2b**	**2c**	**2d**	**2i**	**2j**	**2k**	**2l**	**3a**
Rural	0.003	0.002	0.006**	0.015**	0.015***	0.005+	0.002	0.023***	0.023***	0.005*	0.002	0.013***
	(0.002)	(0.002)	(0.002)	(0.005)	(0.005)	(0.003)	(0.003)	(0.005)	(0.005)	(0.003)	(0.003)	(0.002)
Suburban	0.002	0.005*	0.003	0.026***	0.026***	0.002	−0.002	0.035***	0.036***	0.002	−0.002	0.005**
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.002)	(0.002)	(0.004)	(0.004)	(0.002)	(0.002)	(0.002)
Town	0.006*	0.008*	0.008**	0.004	0.004	0.011**	0.009*	0.009+	0.011+	0.010**	0.009*	0.013***
	(0.003)	(0.003)	(0.003)	(0.006)	(0.006)	(0.004)	(0.004)	(0.006)	(0.006)	(0.004)	(0.004)	(0.002)
Constant	0.069***	0.083***	0.078***	0.139***	0.138***	0.101***	0.103***	0.161***	0.159***	0.102***	0.104***	0.081***
	(0.001)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.001)
***N* schools**	**13,711**	**13,711**	**13,711**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**11,097**
Percentage of students eligible for FRPL	−0.018***	−0.021***	−0.012***	−0.256***	−0.255***	−0.002	0.016***	−0.298***	−0.298***	−0.005	0.015***	−0.010***
	(0.003)	(0.003)	(0.003)	(0.006)	(0.006)	(0.004)	(0.004)	(0.006)	(0.006)	(0.004)	(0.004)	(0.002)
Constant	0.080***	0.096***	0.088***	0.284***	0.283***	0.105***	0.096***	0.334***	0.332***	0.108***	0.097***	0.092***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.002)	(0.002)	(0.004)	(0.004)	(0.002)	(0.002)	(0.001)
***N* schools**	**13,711**	**13,711**	**13,711**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**13,700**	**11,097**

NOTE: + $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, 2014 ed., Washington, D.C.: American Educational Research Association, 2014.

Archbald, Doug, Joseph Glutting, and Xiaoyu Qian, "Getting into Honors or Not: An Analysis of the Relative Influence of Grades, Test Scores, and Race on Track Placement in a Comprehensive High School," *American Secondary Education*, Vol. 37, No. 2, January 2009, pp. 65–81.

Ballotpedia, "School Responses to the Coronavirus (COVID-19) Pandemic During the 2020–2021 Academic Year," webpage, undated. As of January 5, 2021:
https://ballotpedia.org/School_responses_to_the_coronavirus_(COVID-19)_pandemic_during_the_2020-2021_academic_year

Barnum, Matt, "Evidence of Learning Loss Is Piling Up. Here's How the U.S. Could Design a Tutoring Program to Help," Chalkbeat, December 9, 2020. As of January 5, 2021:
https://www.chalkbeat.org/2020/12/9/22165700/learning-loss-tutoring-blueprint-schools

Beaver, Jessica K., and Elliot H. Weinbaum, "State Test Data and School Improvement Efforts," *Educational Policy*, Vol. 29, No. 3, 2013, pp. 478–503.

Bellwether Education Partners, "Data-Driven Back to School Priorities for District and School Leaders," fact sheet, 2020. As of January 5, 2021:
https://bellwethereducation.org/sites/default/files/Bellwether_Accountability-DataDrivenPriorities.pdf

Boyer, Michelle, and Leslie Keng, "Test Score Meaning Under Remote Test Administration," Center for Assessment, October 2, 2020. As of January 22, 2021:
https://www.nciea.org/blog/school-disruptions/test-score-meaning-under-remote-test-administration

Bui, Sa A., Steven G. Craig, and Scott A. Imberman, "Is Gifted Education a Bright Idea? Assessing the Impact of Gifted and Talented Programs on Students," *American Economic Journal: Economic Policy*, Vol. 6, No. 3, August 2014, pp. 30–62.

California State Senate, California Mathematics Placement Act of 2015, SB-359, October 5, 2015. As of January 13, 2021:
https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201520160SB359

California Department of Education, "Academic Performance Calculation: Methodology for Measuring Performance on Academic Performance," webpage, December 30, 2020. As of January 6, 2021:
https://www.cde.ca.gov/ta/ac/cm/acadindcal.asp

Card, David, and Laura Giuliano, "Can Tracking Raise the Test Scores of High-Ability Minority Students?" *American Economic Review*, Vol. 106, No. 10, 2016, pp. 2783–2816.

Center for Research on Education Outcomes, *Estimates of Learning Loss in the 2019–2020 School Year*, Stanford, Calif.: Stanford University, 2020. As of January 5, 2021:
https://credo.stanford.edu/sites/g/files/sbiybj6481/f/short_brief_on_learning_loss_final_v.3.pdf

Chicago Public Schools, "Understanding the Accelerated Placement Act," webpage, undated. As of January 6, 2021:
https://go.cps.edu/accelerated-placement

Conchas, Gilberto, "Structuring Failure and Success: Understanding the Variability in Latino School Engagement," *Harvard Educational Review*, Vol. 71, No. 3, 2001, pp. 475–505.

Cortes, Kalena E., Joshua Goodman, and Takako Nomi, "Intensive Math Instruction and Educational Attainment: Long-Run Impacts of Double-Dose Algebra," *Journal of Human Resources*, Vol. 50, No. 1, 2015, pp. 108–158.

Darling-Hammond, Linda, "Inequality and the Right to Learn: Access to Qualified Teachers in California's Public Schools," *Teachers College Record*, Vol. 106, No. 10, October 2004, pp. 1936–1966.

Data Quality Campaign, Alliance for Excellent Education, Collaborative for Student Success, *Measuring Growth in 2021: What State Leaders Need to Know*, Washington, D.C.: Data Quality Campaign, August 2020. As of January 5, 2021:
https://dataqualitycampaign.org/wp-content/uploads/2020/07/Measuring-Growth-in-2021_What-State-Leaders-Need-to-Know.pdf

Doan, Sy, Jonathan D. Schweig, and Kata Mihaly, "The Consistency of Composite Ratings of Teacher Effectiveness: Evidence from New Mexico," *American Educational Research Journal*, Vol. 56, No. 6, December 2019, pp. 2116–2146.

Domina, Thurston, "The Link Between Middle School Mathematics Course Placement and Achievement," *Child Development*, Vol. 85, No. 5, October 2014, pp. 1948–1964.

Dougherty, Chrys, *Use of Data to Support Teaching and Learning: A Case Study of Two School Districts*, Iowa City, Ia.: ACT, ACT Research Report Series, 2015. As of January 5, 2021:
https://files.eric.ed.gov/fulltext/ED558033.pdf

Douglas, Karen M., and Robert J. Mislevy, "Estimating Classification Accuracy for Complex Decision Rules Based on Multiple Scores," *Journal of Educational and Behavioral Statistics*, Vol. 35, No. 3, June 2010, pp. 280–306.

Ehlert, Mark, Cory Koedel, Eric Parsons, and Michael Podgursky, "Selecting Growth Measures for Use in School Evaluation Systems: Should Proportionality Matter?" *Educational Policy*, Vol. 30, No. 3, 2016, pp. 465–500.

Eissenberg, Thomas E., and Lawrence M. Rudner, *Explaining Test Results to Parents*, Washington, D.C.: ERIC Clearinghouse on Tests Measurement and Evaluation, ERIC Digest No. 102, 1988. As of January 5, 2021:
https://files.eric.ed.gov/fulltext/ED302559.pdf

Favero, Nathan, and Kenneth J. Meier, "Evaluating Urban Public Schools: Parents, Teachers, and State Assessments," *Public Administration Review*, Vol. 73, No. 3, May–June 2013, pp. 401–412.

Figlio, David N., "Names, Expectations and the Black-White Test Score Gap," Cambridge, Mass.: National Bureau of Economic Research, Working Paper 11195, March 2005.

Finkelstein, Neal, Anthony Fong, Juliet Tiffany-Morales, Patrick Shields, and Min Huang, *College Bound in Middle School and High School? How Math Course Sequences Matter*, Sacramento, Calif.: Center for the Future of Teaching and Learning at WestEd, 2012. As of January 5, 2021:
https://www.wested.org/resources/college-bound-in-middle-school-and-high-school-how-math-course-sequences-matter/

Gamoran, Adam, "The Stratification of High School Learning Opportunities," *Sociology of Education*, Vol. 60, No. 3, July 1987, pp. 135–155.

Gamoran, Adam, "Access to Excellence: Assignment to Honors English Classes in the Transition from Middle to High School," *Educational Evaluation and Policy Analysis*, Vol. 14, No. 3, Fall 1992, pp. 185–204.

Gao, Niu, *College Readiness in California: A Look at Rigorous High School Course-Taking*, San Francisco, Calif.: Public Policy Institute of California, 2016. As of January 5, 2021:
https://www.ppic.org/content/pubs/report/R_0716NGR.pdf

Gao, Niu, and Sara Adan, *Math Placement in California's Public Schools*, San Francisco, Calif.: Public Policy Institute of California, 2016. As of January 5, 2021:
https://www.ppic.org/content/pubs/report/R_1116NGR.pdf

Gershenson, Seth, Stephen B. Holt, and Nicholas Papageorge, "Who Believes in Me? The Effect of Student-Teacher Demographic Match on Teacher Expectations," *Economics of Education Review*, Vol. 52, June 2016, pp. 209–224.

Gewertz, Catherine, "It's Official: All States Have Been Excused from Statewide Testing This Year," blog post, Education Week, April 2, 2020. As of January 5, 2021:
https://www.edweek.org/teaching-learning/
its-official-all-states-have-been-excused-from-statewide-testing-this-year/2020/04

Giannini, Stefania, Robert Jenkins, and Jaime Saavedra, "Reopening Schools: When, Where and How?" United Nations Educational, Scientific and Cultural Organization, May 13, 2020. As of January 5, 2021:
https://en.unesco.org/news/reopening-schools-when-where-and-how

Goldhaber, Dan, Malcolm Wolff, and Timothy Daly, "Assessing the Accuracy of Elementary School Test Scores as Predictors of Students' High School Outcomes," Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research, Working Paper No. 235-0520, 2020. As of January 5, 2021:
https://caldercenter.org/publications/
assessing-accuracy-elementary-school-test-scores-predictors-students%E2%80%99-high-school

Gong, Brian, "Fall Educational Assessment: The Information You Need and How to Get It," blog post, Center for Assessment, May 6, 2020. As of January 5, 2021:
https://www.nciea.org/blog/fall-assessment/fall-educational-assessment-information-you-need-and-how-get-it

Grissom, Jason A., and Christopher Redding, "Discretion and Disproportionality: Explaining the Underrepresentation of High-Achieving Students of Color in Gifted Programs," *AERA Open*, Vol. 2, No. 1, January–March 2016, pp. 1–25.

Haertel, Edward H., and Joan I. Herman, "A Historical Perspective on Validity Arguments for Accountability Testing," *Yearbook of the National Society for the Study of Education*, Vol. 104, No. 2, June 2005, pp. 1–34.

Hamilton, Laura, Richard Halverson, Sharnell S. Jackson, Ellen Mandinach, Jonathan A. Supovitz, and Jeffrey C. Wayman, *Using Student Achievement Data to Support Instructional Decision Making*, Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, NCEE 2009-4067, 2009. As of January 13, 2021:
https://ies.ed.gov/ncee/wwc/PracticeGuide/12

Hoel, Paul G., Sidney C. Port, and Charles J. Stone, *Introduction to Probability Theory*, 1st ed., Boston, Mass.: Houghton Mifflin, 1971.

Huang, Chun-Wei, Jason Snipes, and Neal Finkelstein, *Using Assessment Data to Guide Math Course Placement of California Middle School Students*, Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education; and Regional Educational Laboratory at WestEd, REL 2014-040, 2014. As of January 13, 2021:
https://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=345

Huff, Kristen, "National Data Quantifies Impact of COVID Learning Loss; Raises Questions About At-Home Testing," media release, Curriculum Associates, 2020. As of January 5, 2021:
https://www.curriculumassociates.com/-/media/mainsite/files/i-ready/
ca-impact-of-covid-learning-loss-fall-2020.pdf

Jaschik, Scott, and Doug Lederman, *2018 Survey of College and University Admissions Directors: A Study by Inside Higher Ed and Gallup*, Washington, D.C.: Inside Higher Ed, 2018. As of January 5, 2021:
https://www.insidehighered.com/news/survey/2018-surveys-admissions-leaders-pressure-grows

Johnson, Angela, and Megan Kuhfeld, *Fall 2019 to Fall 2020 MAP Growth Attrition Analysis*, Portland, Oreg.: NWEA, 2020. As of January 5, 2021:
https://www.nwea.org/research/publication/fall-2019-to-fall-2020-map-growth-attrition-analysis/

Johnson, Sydney, "Shorter Exams Will Replace California's Annual Smarter Balanced Tests," EdSource, November 5, 2020. As of January 5, 2021:
https://edsource.org/2020/shorter-exams-will-replace-californias-annual-smarter-balanced-tests/643088

Kane, Michael T., "Validation," in Robert L. Brennan, ed., *Educational Measurement*, 4th ed., Westport, Conn.: Praeger Publishers, 2006, pp. 17–64.

———, "Validating the Interpretations and Uses of Test Scores," *Journal of Educational Measurement*, Vol. 50, No. 1, Spring 2013, pp. 1–73.

Kelly, Sean, "The Contours of Tracking in North Carolina," *High School Journal*, Vol. 90, No. 4, April–May 2007, pp. 15–31.

Klein, Alyson, "No Child Left Behind: An Overview," Education Week, April 10, 2015. As of January 5, 2021:
https://www.edweek.org/policy-politics/no-child-left-behind-an-overview/2015/04

———, "The Every Student Succeeds Act: An ESSA Overview," Education Week, March 31, 2016. As of January 5, 2021:
https://www.edweek.org/policy-politics/the-every-student-succeeds-act-an-essa-overview/2016/03

Knoeppel, Robert C., "Resource Adequacy, Equity, and the Right to Learn: Access to High-Quality Teachers in Kentucky," *Journal of Education Finance*, Vol. 32, No. 4, Spring 2007, pp. 422–442.

Kriegler, Shelley, and Thomas Lee, *Using Standardized Test Data as Guidance for Placement into 8th Grade Algebra*, Los Angeles, Calif.: University of California-Los Angeles Math Content for Teachers, 2006.

Kuhfeld, Megan, James Soland, Beth Tarasawa, Angela Johnson, Erik Ruzek, and Jing Liu, "Projecting the Potential Impacts of COVID-19 School Closures on Academic Achievement," *Educational Researcher*, Vol. 49, No. 8, 2020, pp. 549–565.

Kuhfeld, Megan, Beth Tarasawa, Angela Johnson, Erik Ruzek, and Karyn Lewis, *Learning During COVID-19: Initial Findings on Students' Reading and Math Achievement and Growth*, Portland, Oreg.: NWEA, 2020. As of January 5, 2021:
https://www.nwea.org/research/publication/
learning-during-covid-19-initial-findings-on-students-reading-and-math-achievement-and-growth/

Lankford, Hamilton, Susanna Loeb, and James Wyckoff, "Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis," *Educational Evaluation and Policy Analysis*, Vol. 24, No. 1, Spring 2002, pp. 37–62.

LaRoy, Tara, and Lisa Roberts, "9th Grade Registration Class of 2023," presentation slides, Knoxville Public Schools, February 2019. As of January 6, 2021:
https://www.knoxschools.org/cms/lib/TN01917079/Centricity/Domain/3985/
Class%20of%202023%20Registration%20Info.pdf

Livingston, Samuel A., and Charles Lewis, "Estimating the Consistency and Accuracy of Classifications Based on Test Scores," *Journal of Educational Measurement*, Vol. 32, No. 2, Summer 1995, pp. 179–197.

Los Angeles Unified School District, "Middle School Mathematics Pathways Overview," video, March 2017. As of January 6, 2021:
https://achieve.lausd.net/Page/11659

Louis, Karen Seashore, Kenneth Leithwood, Kyla L. Wahlstrom, and Stephen E. Anderson, *Investigating the Links to Improved Student Learning: Final Report of Research Findings*, New York: The Wallace Foundation, 2010. As of January 5, 2021:
https://www.wallacefoundation.org/knowledge-center/Documents/
Investigating-the-Links-to-Improved-Student-Learning.pdf

Marion, Scott, and Chris Domaleski, *The Wyoming Comprehensive Accountability Framework: Phase I*, Cheyenne, Wyo.: Wyoming Select Committee on Statewide Education Accountability, 2012. As of January 19, 2021:
https://www.wyoleg.gov/InterimCommittee/2011/SelectAccountability/
Comprehensive%20Accountability%20Framework%20FINAL_013112.pdf

Marion, Scott, Brian Gong, Will Lorié, and Rebecca Kockler, *Restart and Recovery: Assessment Considerations for Fall 2020*, Washington, D.C.: Council of Chief State School Officers, 2020. As of January 5, 2021:
https://www.ccsso.org/sites/default/files/2020-08/CCSSO_RR_Assessment-Fall-v1(updated).pdf

Martineau, Joseph, Chris Domaleski, Karla Egan, Thanos Patelis, and Nathan Dadey, *Recommendations for Addressing the Impact of Test Administration Interruptions and Irregularities*, Washington, D.C.: Council of Chief State School Officers, 2015. As of January 5, 2021:
https://www.nciea.org/sites/default/files/publications/Computer-Based-Interruptions_110415.pdf

Martinez, José Felipe, Jonathan Schweig, and Pete Goldschmidt, "Approaches for Combining Multiple Measures of Teacher Performance: Reliability, Validity, and Implications for Evaluation Policy," *Educational Evaluation and Policy Analysis*, Vol. 38, No. 4, December 2016, pp. 738–756.

McClain, Mary-Catherine, and Steven Pfeiffer, "Identification of Gifted Students in the United States Today: A Look at State Definitions, Policies, and Practices," *Journal of Applied School Psychology*, Vol. 28, No. 1, 2012, pp. 59–88.

McDonnell, Lorraine M., "Assessment Policy as Persuasion and Regulation," *American Journal of Education*, Vol. 102, No. 4, August 1994, pp. 394–420.

McEachin, Andrew, Thurston Domina, and Andrew Penner, "Heterogeneous Effects of Early Algebra Across California Middle Schools," *Journal of Policy Analysis and Management*, Vol. 39, No. 3, 2020, pp. 772–800.

Means, Barbara, Christine Padilla, and Larry Gallagher, *Use of Education Data at the Local Level: From Accountability to Instructional Improvement*, Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation, and Policy Development, 2010. As of January 13, 2021: https://www2.ed.gov/rschstat/eval/tech/use-of-education-data/index.html

Messick, Samuel, "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning," *American Psychologist*, Vol. 50, No. 9, 1995, pp. 741–749.

Nagin, Daniel S., "Group-Based Trajectory Modeling: An Overview," in Alex R. Piquero and David Weisburd, eds., *Handbook of Quantitative Criminology*, New York: Springer, 2010, pp. 53–67.

National Academies of Sciences, Engineering, and Medicine, *Monitoring Educational Equity*, Washington, D.C.: National Academies Press, 2019. As of January 13, 2021: https://www.nap.edu/catalog/25389/monitoring-educational-equity

National Center for Education Statistics, "2017–18 Common Core of Data (CCD) Universe Files," data files, 2019. As of January 22, 2021: https://nces.ed.gov/ccd/files.asp

National Research Council, *Investigating the Influence of Standards: A Framework for Research in Mathematics, Science, and Technology Education*, Washington, D.C.: The National Academies Press, 2002. As of January 13, 2021: https://www.nap.edu/catalog/10023/investigating-the-influence-of-standards-a-framework-for-research-in

NWEA, MAP Growth, webpage, undated. As of January 22, 2021: https://www.nwea.org/map-growth/

———, "Viewing Linking Studies," webpage, last updated December 23, 2020. As of January 6, 2021: https://www.nwea.org/resource/type/linking-studies/

Oakes, Jeannie, *Keeping Track: How Schools Structure Inequality*, New Haven, Conn.: Yale University Press, 1985.

Olson, Lynn, *The New Testing Landscape: How State Assessments Are Changing Under the Federal Every Student Succeeds Act*, Washington, D.C.: FutureEd, Georgetown University, 2019. As of January 2021: https://www.future-ed.org/wp-content/uploads/2019/09/FutureEdTestingLandscapeReport.pdf

Papageorge, Nicholas W., Seth Gershenson, and Kyung Min Kang, "Teacher Expectations Matter," *Review of Economics and Statistics*, Vol. 102, No. 2, May 2020, pp. 234–251.

Peele, Holly, and Maya Riser-Kositsky, "Map: Coronavirus and School Closures in 2019–2020," Education Week, last updated September 16, 2020. As of January 5, 2021: https://www.edweek.org/leadership/map-coronavirus-and-school-closures-in-2019-2020/2020/03

Pennsylvania Department of Education, *PVAAS Methodologies: Measuring Growth and Projecting Achievement*, Harrisburg, Pa., 2020. As of January 6, 2021: https://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PVAAS/Methodology/PVAASMethodologies.pdf

Peugh, James L., and Craig K. Enders, "Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement," *Review of Educational Research*, Vol. 74, No. 4, Winter 2004, pp. 525–556.

Popham, W. James, "The Fatal Flaw of Educational Assessment," Education Week, March 22, 2016. As of January 5, 2021:
https://www.edweek.org/policy-politics/opinion-the-fatal-flaw-of-educational-assessment/2016/03

RCW—*See* Revised Code of Washington.

Reardon, Sean F., "Educational Opportunity in Early and Middle Childhood: Using Full Population Administrative Data to Study Variation by Place and Age," *Russell Sage Foundation Journal of the Social Sciences*, Vol. 5, No. 2, March 2019, pp. 40–68.

Revised Code of Washington, Title 28A, Common School Provisions; Chapter 28A.320, Provisions Applicable to All Districts; Section 28A.320.195, Academic Acceleration for High School Students—Adoption of Policy. As of January 13, 2021:
https://app.leg.wa.gov/RCW/default.aspx?cite=28A.320.195

Rickles, Jordan H., "Using Interviews to Understand the Assignment Mechanism in a Nonexperimental Study: The Case of Eighth Grade Algebra," *Evaluation Review*, Vol. 35, No. 5, 2011, pp. 490–522.

Rubin, Donald B., *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons, Inc., 1987.

Ryan, Erin, "4 Reasons MAP Growth K–2 Should Be Part of Your Assessment Toolbox," blog post, NWEA, August 29, 2019. As of January 5, 2021:
https://www.nwea.org/blog/2019/4-reasons-map-growth-k-2-assessment-toolbox/

SMMUSD—*See* Santa Monica Malibu–Unified School District.

Santa Monica Malibu–Unified School District, "SMMUSD Math Pathways 6 to 7," email correspondence, May 15, 2020.

SAS, *Technical Documentation of the 2017 TVAAS Analyses*, 2017. As of January 6, 2021:
https://www.tn.gov/content/dam/tn/education/data/tvaas/tvaas_technical_documentation_2017.pdf

Sawchuk, Stephen, "Should Schools Be Giving So Many Failing Grades This Year?" Education Week, December 11, 2020. As of January 5, 2021:
https://www.edweek.org/leadership/should-schools-be-giving-so-many-failing-grades-this-year/2020/12

Schafer, Joseph L., and John W. Graham, "Missing Data: Our View of the State of the Art," *Psychological Methods*, Vol. 7, No. 2, 2002, pp. 147–177.

School Board of Broward County, Florida, *2019–2020 Grades 9–12 Curriculum Guide and Graduation Requirements*, 2019. As of January 6, 2021:
https://www.browardschools.com/cms/lib/FL01803656/Centricity/Domain/20/curriculumguide.pdf

Shepard, Lorrie A., "Chapter 9: Evaluating Test Validity," *Review of Research in Education*, Vol. 19, No. 1, 1993, pp. 405–450.

Stein, Mary Kay, Julia Heath Kaufman, Milan Sherman, and Amy F. Hillen, "Algebra: A Challenge at the Crossroads of Policy and Practice," *Review of Educational Research*, Vol. 81, No. 4, December 2011, pp. 453–492.

Strauss, Valerie, "More Students Than Ever Got F's in First Term of 2020–21 School Year—But Are A–F Grades Fair in a Pandemic?" *Washington Post*, December 6, 2020.

U.S. Department of Education Office for Civil Rights, *Civil Rights Data Collection: Data Snapshot: College and Career Readiness*, Washington, D.C., CRDC Issue Brief No. 3, March 2014.

Wan, Lei, Robert L. Brennan, and Won-Chan Lee, *Estimating Classification Consistency for Complex Assessments*, Iowa City, Ia.: Center for Advanced Studies in Measurement and Assessment, University of Iowa, CASMA Research Report No. 22, 2007. As of January 13, 2021:
https://education.uiowa.edu/sites/education.uiowa.edu/files/documents/centers/casma/publications/casma-research-report-22.pdf

Williamson County Schools, "Best-Fit Math Placement," undated. As of January 6, 2021:
https://tn50000578.schoolwires.net/cms/lib/TN50000578/Centricity/Domain/1254/Math-Placement.pdf

Woods, Julie, *Assessments 101: A Policymaker's Guide to K–12 Assessments*, Denver, Colo.: Education Commission of the States, 2017. As of January 5, 2021:
https://www.ecs.org/wp-content/uploads/Assessments-101_A-policymakers-guide-to-K-12-assessments.pdf