ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-520

A New Framework for Identifying At-Risk Students in Public Schools

Ishtiaque Fazlul
University of Missouri

Cory Koedel
University of Missouri

Eric Parsons
University of Missouri

We develop a new framework for identifying at-risk students in public schools. Our framework has two fundamental advantages *over status quo systems*: (1) it is based on a clear definition of what it means for a student to be at risk and (2) it leverages states' rich administrative data systems to produce more informative risk measures. Our framework is more effective than common alternatives at identifying students who are at risk of low academic performance and we use policy simulations to show that it can be used to target resources toward these students more efficiently. It also offers several other benefits relative to *status quo systems*. We provide an alternative approach to risk measurement that states can use to inform funding, accountability, and other policies, rather than continuing to rely on broad categories tied to the nebulous concept of "disadvantage."

VERSION: January 2022

Suggested citation: Fazlul, Ishtiaque, Cory Koedel, and Eric Parsons. (2022). A New Framework for Identifying At-Risk Students in Public Schools. (EdWorkingPaper: 22-520). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/jzv6-da20

A New Framework for Identifying At-Risk Students in Public Schools

Ishtiaque Fazlul
Cory Koedel
Eric Parsons

January 2022

We develop a new framework for identifying at-risk students in public schools. Our framework has two fundamental advantages over *status quo* systems: (1) it is based on a clear definition of what it means for a student to be at risk and (2) it leverages states' rich administrative data systems to produce more informative risk measures. Our framework is more effective than common alternatives at identifying students who are at risk of low academic performance and we use policy simulations to show that it can be used to target resources toward these students more efficiently. It also offers several other benefits relative to *status quo* systems. We provide an alternative approach to risk measurement that states can use to inform funding, accountability, and other policies, rather than continuing to rely on broad categories tied to the nebulous concept of "disadvantage."

Acknowledgement
We thank the Missouri Department of Elementary and Secondary Education for access to data, Rachel Anderson and Alexandra Ball at Data Quality Campaign for useful comments, and Andrew Estep and Cheng Qian for research support. We gratefully acknowledge financial support from the Walton Family Foundation and CALDER, which is funded by a consortium of foundations (for more information about CALDER funders, see www.caldercenter.org/about-calder). All opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the funders, data providers, or institutions to which the author(s) are affiliated. All errors are our own.

1. Introduction

There have been substantial advances in education data infrastructure since the turn of the 21st century and as of our writing this article, virtually every state in the U.S. has a state longitudinal data system (SLDS) supported by large investments from the federal government.[1] These data systems allow states to track students as they move through K-12 schools, monitoring their academic progress and providing rich information about their circumstances. Computing power has also increased rapidly during this same period of investment in data infrastructure so not only are rich data on K-12 students increasingly available, they are also increasingly usable.

However, these gains in data availability and useability have not translated into meaningful improvements in how states identify students in need of additional resources and supports, who are commonly described as being "disadvantaged" or "at risk." States ubiquitously rely on categorical indicators associated with the concept of disadvantage, broadly defined, to identify these students. For example, states have historically used free and reduced-price meal (FRM) eligibility to identify high-poverty students, with some states shifting more recently to the use of direct certification (DC) status. States also group students by characteristics such as English language learner (ELL) status, individualized education program (IEP) status, and underrepresented minority (URM) status, among others. States use combinations of these categorical indicators in a variety of policies, most notably in funding formulas and to track achievement gaps.

The category-based approach used by states is reasonable but lacks a firm conceptual grounding. It is also antiquated from an analytic perspective. With respect to the former critique, it is useful to ask why states use the categories they use to identify at-risk students. A sensible answer is that data exist and the categories correlate with challenging student circumstances. But the precise dimensions of risk targeted by the categories are not clear. For example, consider a definition of risk based on the likelihood of poor academic performance, which we argue below is an appropriate definition in the context of the education system. By this definition, surely there are many FRM-eligible students who are not high risk, and many non-FRM-eligible students who are high risk. Clarifying what we mean by "risk"—that is, risk of what?— illuminates the inaccuracy of common categorical indicators. The inaccuracy is not surprising because these

[1] New Mexico is the lone exception (see here: https://nces.ed.gov/programs/slds/stateinfo.asp, information retrieved 08.23.2021).

indicators are blunt descriptors of student circumstances, but taken at face value, state systems that track students categorically do not acknowledge this problem.

Our other critique of the current approach is that does not leverage states' rich administrative data to improve measurement accuracy. For example, if FRM status is a risk indicator and ELL status is a risk indicator, what about a student who is both FRM-eligible and an ELL? Relatedly, what about a student who is consistently FRM-eligible versus one who is eligible for just a single year? In fact, in the latter case it has been documented empirically that students who are FRM-eligible for multiple years are at greater risk of poor academic performance (Michelmore and Dynarski, 2017). However, state systems do not allow for this type of differentiation to impact students' risk designations. These examples illustrate the general point that existing categorical systems do not leverage the full breadth of information available about students.

With the limitations of existing systems as motivating context, we develop and test a new framework for identifying at-risk students in public schools. Our framework includes a clear definition of risk based on academic performance. It also uses simple but modern methodological tools to better leverage information available in state data systems to measure risk.

We implement our framework using the Missouri SLDS as a proof-of-concept exercise and highlight two key findings. First, our framework is more effective than *status quo* systems at identifying students at risk of poor academic outcomes. This is by design and in a policy context, it means the metrics that emerge from our framework can be used to better target resources toward low-performing students and their schools. The second key finding is that the metrics from our framework can also be used to improve the targeting of resources to students across a broad range of traditional "categories of disadvantage"—namely, ELL, IEP, and URM students—compared to hypothetical systems based on poverty proxies (i.e., FRM and DC status) or a system modeled after California's progressive Local Control Funding Formula (LCFF).

Our framework provides a principled structure within which states can work to develop better measures of student risk. That said, it is not a panacea and has limitations that we elaborate on over the course of this article. Some of the limitations are inherent to the challenging problem of measuring student risk, but we expect most can be improved upon with additional research. Our goal is to propel research forward on a more promising path toward the accurate and useful measurement of student risk. In turn, this can improve the efficacy of state policies designed to

narrow achievement gaps and promote better academic outcomes among at-risk students.

2. Why a New Framework is Needed

In this section we expand on the two fundamental features of our framework that distinguish it from *status quo* systems, which are: (1) it establishes a clear definition of student risk (based on academic performance) and (2) it applies modern data and analytic tools to measure risk more effectively.

First, we assert that a clear definition of risk is a desirable feature of any framework used to identify at-risk students. Once a definition is established, a framework's efficacy can be assessed by the accuracy with which at-risk students are identified. *Status quo* systems, while aimed in the general direction of identifying students in need of supports, do not clearly define what it means for a student to be at risk. This makes it impossible to assess their efficacy. For example, even for a concept as straightforward as poverty, we are not aware of any state system that formally acknowledges the difference between available measures of poverty used for policy purposes—which are either inaccurate, heavily coarsened, or both—and actual poverty.

The process of defining "risk" in the abstract is undoubtedly challenging. However, we argue that in the context of schooling risk can be defined based on academic performance. In our framework, we define a student as at risk if her characteristics indicate she is likely to perform poorly in school.

This is an intuitive and reasonable definition, but we acknowledge alternative definitions exist, of which perhaps the most credible is a poverty-based definition. However, even if one concedes that definitions of risk based on academic performance and poverty both have conceptual merit, academic performance is preferable for two practical reasons. The first reason is that we are currently unable to measure poverty accurately in the education system. This is an uncomfortable truth, but an important one. To illustrate this point, we briefly review the prospects for the two predominant poverty metrics used in state policies: (1) FRM eligibility from the National School Lunch Program (NSLP), and (2) DC status based (primarily) on participation in social safety net programs outside of schools. Regarding FRM data, Domina et al. (2018) and Fazlul, Koedel, and Parsons (2021) show conclusively that these data do not measure poverty accurately. Fazlul, Koedel, and Parsons (2021) find that DC data measure poverty accurately in Missouri, at least on average at the threshold of 130 percent of the poverty line. However, DC status is limited because it is a heavily-coarsened measure of poverty.

Moreover, cross-state variability in direct certification processes is such that DC status will be more effective in some states than others at identifying high-poverty students.[2] If the goal is to measure student poverty, it must be acknowledged that education systems do not have access to accurate, consistent, and differentiated measures of poverty.[3]

The second practical problem with poverty measurement in education is that states rely entirely on data from external programs. Historically, FRM data have come from the U.S. Department of Agriculture's NSLP and the primary program used to determine DC status is the Supplemental Nutrition Assistance Program (SNAP). This is problematic because external programs can change. A recent example is the introduction of the Community Eligibility Provision (CEP) to the NSLP, which allows sufficiently high-poverty schools to offer free meals to all of their students. It was not the intent of the CEP to alter the informational content about poverty contained by FRM status, but in many states, this is what happened (Chingos, 2018; Greenberg, 2018; Koedel and Parsons, 2021). Now states are turning to DC data to replace FRM data but the fundamental limitation remains: changes to eligibility requirements of the core programs that determine DC status will change the informational content of the data. A desirable feature of a stable measurement system in education is that to the extent possible, student designations should not be subject to changes caused by changes to external programs.

The second core consideration that motivates our framework is the failure of current systems to leverage all of the information available about students to assess their risk levels. This consideration is independent of how risk is defined. The use of blunt categories in current systems is only preferable if there is no marginal information to be extracted from additional variables in state data systems and no degrees of differentiated risk indicated by the persistence of students' categorical assignments or belonging to multiple categories. This condition is intuitively implausible and has been refuted empirically for specific variables in recent studies by Goldhaber et al. (2022) and Michelmore and Dynarksi (2017).

[2] To elaborate briefly, different rules for direct certification across states are such that the poverty threshold identified by DC data will vary. In some cases, it may be difficult to tie DC status to a particular threshold. Another issue is that direct certification relies on families' participation in social safety net programs and not all high-poverty populations are similarly likely to participate, with Hispanics being a prime example of a group underrepresented in programs that lead to direct certification. See Fazlul, Koedel, and Parsons (2021) for further discussion.

[3] Fazlul, Koedel, and Parsons (2021) also evaluate the accuracy of School Neighborhood Poverty (SNP) data from the National Center for Education Statistics. They show that SNP data measure poverty accurately on average at the school level, but there are many sources of errors for individual schools and no SNP metrics are available for individual students.

3. A New Framework for Measuring Student Risk Status in K-12 Public Schools

3.1 The Framework

We develop our framework for measuring student risk around the most reliable, scalable indicators of academic performance available in the education system: state standardized assessments. Although other informative measures of academic performance surely exist in pockets of the education system—e.g., well-designed tests and assignments in particular teachers' classrooms, personal information about individual student circumstances known by individual educators, etc.—no better and more differentiated information about student academic performance exists at scale. A plausible alternative would be a marker of academic progress, such as high school graduation or college matriculation, but key weaknesses of these metrics include (1) they are observed less often (e.g., only once at th e point of high school exit for each student) and (2) they are not as differentiated as test scores (Austin et al., 2021). State accountability policies also already emphasize standardized assessments, making them a natural focal point for measuring academic performance. Moreover, research causally links student test scores to consequential later life outcomes such as college attendance and earnings.[4]

The foundation of our framework is a predictive linear regression of student test scores using student attributes, which can be expressed generically as follows:

$$S_i = \beta_0 + \mathbf{X_i}\boldsymbol{\beta_1} + \varepsilon_i \quad (1)$$

In equation (1), S_i is a test score for student *i* and $\mathbf{X_i}$ is a vector of student attributes. For the moment, $\mathbf{X_i}$ can be thought of as capturing information about students along a variety of dimensions and of a variety of types (e.g., contemporary and historical information, individual and school-level information, interactions of individual attributes within the vector, etc.). We will be more precise about how we specify $\mathbf{X_i}$ in our proof-of-concept application using the Missouri SLDS below.

The predicted values from this regression, $\hat{S}_i$, can be interpreted as measures of student risk. They are weighted averages of the attributes in the vector $\mathbf{X_i}$, where the weights—the coefficients in the vector $\boldsymbol{\beta_1}$—depend on the extent to which each attribute predicts student

[4] For a recent review of research and discussion on this point see Goldhaber and Özek (2019).

performance. Students with lower values of $\hat{S}_i$ are at greater risk of poor academic performance than their peers with higher values, as determined by their attributes.

An immediate question is the following: If the aim is to define risk status based on test performance, why bother estimating $\hat{S}_i$ when the actual test score, S_i, is observed? There are two reasons, one practical and one conceptual. The practical reason is that S_i is only available for test takers, but values of $\hat{S}_i$ can be calculated for all students regardless of whether they are tested. That is, we can use the output from equation (1) to extrapolate from the test-taking population to produce common measures of risk for all students, inclusive of students who do not take the test, using their attributes in the $\mathbf{X}$ vector and the weights $\boldsymbol{\beta}_1$. The extrapolation to untested students (including those outside of tested grades) requires assuming the attributes that predict performance for tested students are the same attributes that would predict performance for untested students, had they been tested. Although we cannot test this assumption directly because scores for untested students are unobserved, it is intuitive. We also provide evidence consistent with this assumption being upheld, at least to an approximation, by showing that we obtain similar values of $\hat{S}_i$ for individual students when we estimate equation (1) using different subsamples of tested grades (see below).

The conceptual reason for using $\hat{S}_i$ instead of S_i is that it creates a profile-based prediction of student performance that does not depend on the student's actual performance. To understand why this is appealing we must consider the intended use of the risk metrics we aim to develop, which is to inform consequential state policies. We have two types of policies in mind: funding policies and accountability policies.

Given this intended use, it is a general principle of our framework that the measures themselves should be impervious to the activities of educational actors (i.e., districts, schools, and teachers) to the extent possible. To illustrate with a counterexample, consider a system where funding increases are provided to support at-risk students and risk status is defined by observed test scores, S_i. This would perversely incentivize educational actors to produce lower test scores and, in doing so, would harm the credibility of the system.[5] However, if risk status is

[5] Note that whether districts, schools, and teachers actually respond to their perverse incentives has no bearing on the desirability of this feature. For example, even if no actor responded to the incentive to produce low test scores,

defined by $\hat{S}_i$—that is, by how students are predicted to perform based on their attributes given statewide performance patterns—and if the underlying attributes cannot be manipulated by educational actors, this undesirable design feature disappears.

More accurately, it *nearly* disappears, because even if the elements of **X** used to predict test scores are selected to be non-manipulable, the weights contained by $\boldsymbol{\beta}_1$ may still be influenced by school demographics and policies. A concern would be if a particular district enrolls a disproportionate share of students with one of the predictive attributes, say X_k. In that case, the district's own performance could meaningfully influence the weight on that attribute, β_k. Fortunately, this problem can be overcome through jackknifing, which is an estimation procedure whereby the weights applied to produce the values of $\hat{S}_i$ for students in a particular school or district are based only on data from outside that school or district. With the use of jackknifed estimates of $\boldsymbol{\beta}_1$, and elements of **X** that cannot be influenced by educational actors (more on this below), $\hat{S}_i$ is a non-manipulable indicator of student risk that can be applied in consequential education policies.

3.2 Assessing the Framework via Policy Simulation

We present standard statistical diagnostics for the prediction model that underlies our framework in the empirical application below, but our real examination of the value of our framework is based on a policy simulation. The simulation illustrates how our metrics would affect the allocation of resources to different types of students and the schools that serve them. We begin by using the $\hat{S}_i$ values for individual students to produce new a binary categorization of risk status based on predicted academic performance. To do this, we set a threshold value $\tilde{S}$. If $S_i < \tilde{S}$ we assign student *i* as "high risk," otherwise we assign the student as "low risk." These binary categories replicate the categorical (and mostly binary) structure of other risk-status measures—e.g., DC, ELL, FRM, IEP, URM, etc.—with the key difference being that they separate students based on their risk of low academic performance. The binary categories discard

schools and districts that did in fact produce them would receive more funding, giving the impression that poor performance is rewarded and undermining the system.

valuable information about differentiated student risk by coarsening $\hat{S}_i$, but are useful because they facilitate apples-to-apples comparisons to *status quo* systems.[6]

Our policy simulation is based on the following progressive formula for allocating resources to "high risk" and "low risk" students:

$$N_L + (1+Z)N_H = B \qquad (2)$$

In equation (2), N_L is the number of low-risk students, N_H is the number of high-risk students, and *B* is the total resource budget. The amount allocated to each low-risk student is normalized to 1.0 and *Z* is a positive multiplier that allows more per-pupil resources to be distributed to high-risk students. N_L and N_H are choice variables that depend on how low-risk and high-risk students are defined. We can use $\hat{S}_i$ to assign students to low-risk and high-risk categories, or we can assign students using traditional categories such as FRM status, DC status, ELL status, IEP status, and URM status.

A simple way to think about the total resources denoted by *B* is in dollar terms, but the framework is broader than that. For example, *B* could reflect the availability of a specific, centrally-allocated resource like additional personnel or tutoring services within a state. Generally speaking, equation (2) can be thought of as describing the allocation of any scarce resource across low-risk and high-risk students in the system.

The values of N_L and N_H, determined by the definitions of "low risk" and "high risk" students, along with the fixed budget *B*, will yield different values of *Z* as described by the following re-arrangement of equation (2):

$$Z = \frac{B - N_L}{N_H} - 1 \qquad (3)$$

We impose the constraint that $B > N = N_L + N_H$, which ensures that *Z* is positive. In other words, this constraint ensures there is enough funding to provide more than one normalized resource unit for each high-risk student.

Policy simulations based on this resource-allocation model allow us to document how different measures of risk status yield different resource allocations across students and schools

[6] Another advantage of our framework is that more differentiated information about student risk can be recovered from the uncoarsened values of $\hat{S}_i$. See below for further discussion.

with different characteristics. While we use this resource-allocation model as our primary policy simulation, we also briefly consider an accountability policy application as a separate extension.

4. Data and Implementation

4.1 Data Overview

We use administrative microdata from the Missouri SLDS for our proof-of-concept empirical application. The Missouri SLDS is typical of other state systems nationwide; therefore, the application should generalize broadly to other states. Of course, different student and circumstantial attributes may be differentially predictive of student performance in different states, but the structure of the framework should generalize.

The foundation of our framework is the cross-sectional regression described in broad terms by equation (1), and accordingly, we conduct our analysis using the Missouri SLDS with just one cohort of students from the 2016-17 school year (hereafter: 2017). That said, some of the student variables we use to predict test scores are longitudinal, such as persistent poverty. We construct these variables by looking backward in the SLDS for students in the 2017 cohort, although we still estimate the prediction regressions cross-sectionally.

There are two major components of our prediction framework: outcomes and predictors. For outcomes, our primary specification is based on student test scores on state assessments in math and ELA in grades 3-8. We standardize each test by subject-grade and define S_i for each student as the average standardized score across subjects. We restrict the analytic sample we use to estimate equation (1) to students with test scores in both subjects. Recall that this does not influence our ability to produce estimates of $\hat{S}_i$ for all students because we assume the predictors of academic performance are the same for tested and untested students. Under this assumption, we produce values of $\hat{S}_i$ for all students in Missouri in grades 3-12.[7]

We cannot test the assumption that the predictors are the same for tested and untested students directly. However, to gain some insight into its plausibility, we test whether students' predicted scores, $\hat{S}_i$, are sensitive to using different combinations of tested grades to estimate equation (1). For instance, if we obtain substantively different values for $\hat{S}_i$ depending on

[7] Values can also be assigned to students in grades K-2 using the same procedure. There are some technical implications with respect to variable construction for younger students that merit consideration—namely for the panel variables we describe below given that students' data histories do not begin until kindergarten—but in principle the predictions can be extended to earlier grades.

whether we estimate the model on students in grades 3-8, versus grades 3-5, versus grades 6-8, it would suggest the predictors are grade-level sensitive. But if we obtain similar values of $\hat{S}_i$, it would suggest the predictors are generally stable across tested groups. In Appendix Table A1, we show that values of $\hat{S}_i$ estimated using different subsets of tested grades are highly correlated.

There are two ways states could specify the threshold test value, $\tilde{S}$, separating low-risk and high-risk students. One way is to pre-specify a fraction of students the state wishes to target for additional resources. Equation (3) gives a rationale for this approach by showing the fraction of students identified determines the level of additional resources available per high-risk student under a fixed budget. Identifying too many high-risk students dilutes the resource level per student. Another way to set $\tilde{S}$ is to anchor it to a proficiency-category threshold on the state test, which is appealing in that it ties the framework directly to state education policy objectives.

We build our policy simulation following the proficiency-based approach, although we implement a simplified version. The simplification is that we set a single threshold value of $\tilde{S}$ for all grades based on 2017 Missouri NAEP performance. Averaging across math and English Language Arts in grades 4 and 8, NAEP data show that 26.25 percent of Missouri students score below basic. We use this percentile threshold—the 26.25th percentile—as $\tilde{S}$ and assign all students to high-risk and low-risk categories based on their predicted scores on the Missouri state test (the Missouri Assessment Program, or MAP). That is, we assign students with values of $\hat{S}_i$ below the 26.25th percentile as high-risk students, and students above the 26.25th percentile as low-risk students.

Our simplified approach to setting the threshold value of $\tilde{S}$ based on NAEP data does not have any substantive bearing on how our framework operates. It is appealing because it simplifies the analysis and improves the generalizability of our findings by creating a degree of separation from the specific policy context in Missouri. A state wishing to implement our framework with closer adherence to its own grade-and-subject-specific standards could set the $\tilde{S}$ values separately for each grade (and even subject); this would not affect the efficacy of the framework in any substantive way.

That said, an important feature of $\tilde{S}$ is that it is percentile-based rather than based on a raw test score value. This is necessary because the predicted scores, $\hat{S}_i$, are implicitly shrunken

through the prediction process and as a result, the distribution of $\hat{S}_i$ is tighter than the distribution of S_i. The use of a score-based value to set $\tilde{S}$ would result in a lower share of high-risk students identified than students whose actual test scores are below the threshold value.[8]

Next, we turn to the predictors of academic performance. We consider three broad types of predictor variables: (1) individual-level contemporaneous variables, (2) individual-level panel variables, or persistence variables, and (3) school-average variables. The individual variables are listed in Table 1 and include measures of student mobility (number of districts attended in year *t*, number of schools attended in year *t*), ELL status, IEP status, race-ethnicity category (where the categories are American Indian, Asian/Pacific Islander, Black, Hispanic, White, and Multi-race), gender category (male or female), FRM status, and DC status. The panel variables are three-year averages of the individual-student variables taken over the current and two preceding years.[9,10] These variables capture the persistence of students' circumstances, motivated by prior work on the predictive validity over test scores of persistent poverty (Michelmore and Dynarski, 2017) and mobility (Goldhaber et al., 2022). The third and final set of variables includes school averages of the contemporaneous student variables. The school-average variables capture the predictive influence of schooling circumstances conditional on individual student circumstances.

4.2 Practical Issues Associated with Taking the Framework to the Data

4.2.1 Variable Selection

In Section 3 we discuss how in the ideal implementation of our framework, the variables included in the **X**-vector would be non-manipulable. The non-manipulability principle led us to

[8] Alternatively, a variance inflation procedure like the one discussed in Appendix E could be used to set score-based thresholds.

[9] We exclude variables that generally do not change over time, such as race-ethnicity and gender designations, from the panel variable list. For the mobility variables, we divide the total numbers of schools and districts attended by the number of years the student was enrolled in a Missouri school district in the last three years, then additionally control for the fraction of years the student was enrolled in a Missouri district. This three-variable set captures mobility between Missouri schools and districts and across state lines over the three-year period.

[10] The use of three-year averages rather than, say, count variables, allows us to use three years of data for students who we observe for at least three years and fewer years for students new to Missouri or with missing data. For example, a student with two years of data who is FRM-eligible in both years will be coded as "100 percent" FRM-eligible for the panel variable and similarly for a student with just one year of data. The alternative is to use count variables and drop students with insufficient histories (who analytically will be treated as having incomplete **X** vectors). However, this is an inferior option from a policy perspective because it would lead to fewer students for whom we can estimate $\hat{S}_i$. The efficacy, or lack thereof, of our panel-variable models embodies both the substantive importance of the variables and the limitations associated with using them in a comprehensive way as would be necessary in a real policy application.

exclude some types of information available in the SLDS from the predictor set at the onset—examples include data on student attendance, behavioral incidents, course-taking, and grades. These and related variables are typically included in SLDS-based "early warning systems" designed to identify students at risk of poor academic outcomes, such as high school dropout (Li et al., 2016). However, although these kinds of variables are predictive, they are a poor fit in our framework because they can be affected (potentially a great deal) by district and school behavior. Per the preceding discussion, their inclusion would create perverse incentives in the policy applications we have in mind, which have high-stakes funding and accountability consequences attached.[11]

While we omit some of the most manipulable variables from the prediction framework, we also acknowledge that not all of the remaining variables listed in Table 1 are entirely non-manipulable. For example, schools and districts can manipulate FRM status by adopting community eligibility, if eligible; and if not, they can manipulate individual student designations through other aspects of the NSLP application process (Bass, 2010). Schools and districts can also potentially manipulate other student categories including ELL and IEP. Unfortunately, there are few strong predictors of student performance in the Missouri SLDS that are entirely non-manipulable, which suggests a tradeoff between the predictive validity of $\hat{S}_i$ and its manipulability. We were unable to construct a credible version of $\hat{S}_i$ that excludes all potentially manipulable student attributes. Instead, we take a middle-of-the-road approach by allowing some potentially manipulable variables into the framework. The tradeoff between non-manipulability and predictive accuracy merits consideration from users of our framework in any policy application.

Ultimately, our preferred model includes all of the variables listed in Table 1, but uses DC status in place of FRM status as the measure of student poverty. We favor the use of DC status over FRM status because it is a more accurate measure and cannot be manipulated as easily as FRM status (Fazlul, Koedel, and Parsons, 2021). After making this switch, students' ELL and IEP designations are the most manipulable categories remaining in our preferred implementation of our framework.

[11] See Public Impact with Education Analytics (2021) for a related application. Their initial framework does not account for non-manipulability, but they emphasize its importance in their discussion of key issues for future research.

An additional variable-selection issue highlighted by this discussion is with regard to the use of information external to the education system. An application of our framework that is entirely internal to the education system would exclude both FRM and DC data (because both derive from external programs), but as we show below, this has consequences for the predictive accuracy of equation (1). Our decision to include DC status as an external indicator of poverty is again in the spirit of striking a balance between competing priorities. This choice is made easier by evidence we show below that students' risk designations based on our framework are far less sensitive to data disruptions (e.g., a policy change that makes a variable unavailable or changes its meaning) than risk designations in common categorical systems.

4.2.2 Variable Weights

In addition to concerns about the manipulability of variables in the **X**-vector, we must also be concerned about the variable weights ($\boldsymbol{\beta}_1$ in equation 1). Districts and schools can influence these weights if they serve a disproportionately large fraction of students with particular attributes. For example, consider a case of extreme residential segregation by race-ethnicity in a system with two districts, A and B. If District A predominantly serves URM students and is also highly effective, the race-variable weights in equation (1) will partly reflect District A's effectiveness, leading to lower "risk" scores for URM students than would be implied by non-schooling conditions alone.

Jackknifing is an estimation procedure that prevents individual schools and districts from influencing their own weights. In our application, it involves estimating multiple iterations of equation (1) after removing some data at each iteration. In its purest form, a district-level jackknife with *J* districts involves estimating *J* "leave-one-out" versions of equation (1), where each version is estimated on *J-1* districts.[12] The version estimated for individual district *j* includes data from all districts except *j* itself. The jackknifed fitted values for district *j* are a function of the characteristics of students in district *j* (which are unchanged by the jackknifing procedure), **X**, and a set of weights, $\boldsymbol{\beta}_1^j$, unique to district *j* and estimated using data entirely outside of district *j*. Conceptually, these fitted values can be described as capturing the degree of risk of students in district *j* based on their attributes, as predicted by a statewide model outside of

[12] We jackknife at the district level throughout our application. Jackknifing at the school level is also possible, but it is less conservative, more computationally intensive, and unnecessary because district jackknifing works well empirically (see below).

district *j*. The jackknifed estimates of the weighting parameters have the desirable feature that they cannot be influenced by district *j*'s own behavior.

Jackknifing is a common procedure used in academic research, but at least in its purest form, it can be computationally intensive and may be unnecessarily complex for policy applications such as the one we consider. Therefore, we explore the use of simpler variants of the jackknifing procedure. Our preferred jackknife is what we refer to as a "random-quarters" jackknife, which randomly divides districts in Missouri into four equal-sized groups and estimates four "leave-one-group-out" jackknifed versions of equation (1). Each district's jackknifed values are from the regression that excludes the random quarter of the sample to which it belongs. In the appendix we confirm that other jackknifing approaches yield similar results—e.g., splitting the sample randomly into thirds, fifths, tenths, and a full jackknife (see Appendix Table A2). All of the results presented below use the random-quarters jackknife.

5. Empirical Application

5.1 Descriptive Documentation of Model Output

Table 2 provides statistical summary information for variants of equation (1) that include different combinations of variables in **X**. Rows (*a*)-(*d*) include only student-level contemporaneous variables to predict test scores, rows (*e*)-(*h*) build on the models in rows (*a*)-(*d*) by adding corresponding panel variables, and rows (*i*)-(*l*) further add corresponding school-level variables. Within each set of rows, the models become increasingly rich moving down in the table. The last row within each horizontal panel (rows (*d*), (*h*), and (*l*)) also includes two-way interactions of all variables for the relevant variable types (the types are: individual student, panel, and school-level). The notes to Table 2 give precise details about each specification.

The columns of the table provide statistical information about the models. In column (1), the R-squared values range from about 0.21 in our sparsest specification to 0.29 in the models with the most predictive power. Our preferred specification is shown in row (*l*), where the R-squared is at the maximum in the table. Row (*l*) uses all available information, includes two-way interactions between the individual student, panel, and school-aggregate variables, and uses DC status instead of FRM status to capture economic disadvantage.

Note the maximum possible R-squared value for each specification in Table 2 is below 1.0. This is because there is test measurement error in S_i and school effects explain some of the variance in student outcomes. However, variation from these sources is not predicted by the

variables in **X**. This puts a ceiling on the maximum feasible R-squared value in Table 2; a rough estimate is that the maximum should fall in the range of 0.70-0.80.[13] Scaling the estimated R-squared from our preferred specification in row (*l*) by the center of this range—0.75—gives an *ad hoc* "effective R-squared" of 0.39.[14] This is our best estimate of the share of the explainable variation in student test scores accounted for by our preferred model.

Unfortunately, it is difficult to gain insight from this number about the efficacy of our predictions. An R-squared value that is too low is undesirable because it would imply poor predictions from the model, but an R-squared that is too high is undesirable because some distance between S_i and $\hat{S}_i$ is appealing from an incentive-design perspective, per the preceding discussion. The R-squared values reported in Table 2 do not seem particularly "high" or "low" at a cursory glance, although it is a diagnostic limitation that there is no concrete way to judge the performance of the model in this way.

Complementing the R-squared values, Column (2) shows MSEs for the individual predictions relative to observed test scores, and columns (3)-(5) show error rates for the binary predictors of which students are below the score threshold. For the results in these latter columns, we assign the lowest 26.25 percent of students based on $\hat{S}_i$ to the below-basic category then compare their predicted assignments to their actual assignments based on being below the 26.25th percentile in the distribution of S_i. A false-positive is a student we assign as "high risk" based on $\hat{S}_i$, but who scores at or above the 26.25th percentile in reality; and vice-versa for a false negative. The MSE and error-rate numbers come with the same interpretive caveats as the R-squared values: numbers that are too high, or too low, are both of concern.

Although it is difficult to draw conclusions about the general efficacy of our framework from Table 2, it is useful for comparing the different specifications of equation (1). One takeaway from the table is that poverty status data—whether FRM or DC data—add substantial

[13] First, measurement error attributable to the testing instruments accounts for about 10 percent of the variance in these tests (e.g., see Data Recognition Corporation, 2019), and following Boyd et al. (2013), if we use a broader definition of test measurement error it roughly doubles this value to 20 percent. In addition, based on Konstantopoulos and Borman (2011), unobserved factors across schools—inclusive of (and arguably primarily consisting of) school effects—can be estimated to account for up to an additional 10 percent the variance in scores. Subtracting these variance shares from the maximum R-squared value of 1.0 yields a feasible maximum in our application in the range of 0.70-0.80.

[14] In instances where the upper bound R-squared value is below 1.0, the effective R-squared can be obtained by dividing the estimated R-squared by the maximum value (e.g., Aaronson, Barrow, and Sander, 2007).

predictive value to the model. Relative to our specifications in rows (*a*), (*e*), and (*i*) that omit this information, the R-squared increases by about 3-5 percentage points in rows where we include it in various forms. Between the two, FRM data are more predictive of test scores than DC data (this can be seen by comparing the output in rows (*b*) and (*c*), (*f*) and (*g*), and (*j*) and (*k*)). In results omitted for brevity, we confirm that DC status is a stronger predictor of low test scores than FRM status for individual students. However, DC data contribute less to the explanatory power of the model because there is more variance in FRM data (i.e., the FRM-eligible student share is closer to 0.50).[15]

Another takeaway from the table is that conditional on the first-order variables, there is only a marginal gain in explanatory power from adding the interaction variables to the models. This can be seen by the small changes in the R-squared values, MSEs, and error rates corresponding to the rows in Table 2 that add the interaction terms ((*d*), (*h*), and (*l*)) relative to their preceding rows. The limited impact of the interaction variables does not mean that student assignments to multiple categories do not matter—the models without interactions still allow students who belong to multiple categories to have lower predicted performance. Rather, the limited impact of the interactions suggests that the predictive influence of multi-category assignment can be inferred (roughly) additively. For this reason, we do not pursue more complex models with additional interactions.[16]

Table 3 summarizes our risk measures overall, and within traditional categories of disadvantage, by reporting means and standard deviations of $\hat{S}_i$ from our preferred specification in Table 2, row (*l*).[17] First, focusing on the group-average values of $\hat{S}_i$ in the first row of Table 3, the results reflect the well-understood achievement gaps that motivate the policy focus on these categories. The gaps in average predicted achievement by DC status, FRM status, ELL status,

[15] Recall from above that our preference for using DC data is not based on maximizing predictive power, although it is helpful that there is not a major loss of predictive power in switching from FRM to DC data, especially in our richest specifications. Below we show that the predicted values, $\hat{S}_i$, are very highly correlated in models that switch between using FRM and DC data.

[16] Future work could apply machine learning tools to select the optimal model. However, while this would be an interesting academic exercise, our investigation suggests that meaningful gains in predictive power are unlikely. It is also a policy consideration that the use of machine learning could be perceived as less transparent. We leave consideration of the costs and benefits of alternative estimation techniques within our framework to future research.

[17] None of the substantive findings in Table 3 are unique to using our preferred specification.

URM status, and IEP status are 0.58, 0.52, 0.44, 0.59, and 0.94, respectively. These gaps are in standard deviation units of test scores and large by any reasonable standard.[18]

It is a useful (albeit predictable) validity check of our framework that it replicates well-established achievement gaps on average between the categories in Table 3. But the more important new information is in the second row of the table, which reveals broad heterogeneity in the risk for poor academic performance *within* traditional categories of disadvantage. To see this, first note that column (1) shows that across all students, the standard deviation of $\hat{S}_i$ is 0.50.[19] The subsequent columns show there is almost as much variation in $\hat{S}_i$ within several of the categories as in the full sample—e.g., the standard deviations within the FRM-eligible category, non-ELL category, and URM category are all 0.49. Table 3 provides empirical support for the intuitive claim that traditional categories of disadvantage used in state policies are coarse and mask considerable variability in student risk as measured by academic performance.

While Tables 2 and 3 provide necessary contextual information, they are not directly informative about the utility of our framework. This is because they do not address the policy-relevant question of whether our risk measures are "good enough to be useful." The answer to this question depends on the policy objective and the quality of alternative options. In the next section we incorporate these dimensions via our policy simulations.

5.2 Policy Simulations

Table 4 shows results from our first set of policy simulations. We use the risk measures from our framework to allocate resources to students and compare the allocations to alternative allocations using the same policy but where risk status is defined by DC or FRM status. We set $B = 1.25N$ (recall N is the total number of students). Our results are not directionally sensitive to the value of *B,* but all else equal, larger values of *B* generate larger resource gaps between high-risk and low-risk students.

[18] We do not report values for the many coefficients from our prediction models because the multivariate regression framework makes the interpretation of individual coefficients intractable, especially in our richer (and preferred) specifications. That said, the mean values of $\hat{S}_i$ across student categories in Table 3 permit inference about the net direction of the model predictions. More information about the performance of the prediction model can be found in Appendix B.

[19] This value is below 1.0 due to shrinkage in the predictions. Table 1 (including the table notes) shows that the raw standardized scores have standard deviations of approximately 1.0, which is by construction.

Each column of Table 4 shows results from a different policy parameterization, defined by the first four rows. The rows in the lower panel of the table show the average resource units accruing to students with different characteristics. It is these rows that show the policy impacts of our framework, in the form of changes to the resource allocations compared to DC- and FRM-based alternatives.

We walk through how to read the table using the results in column (1) under the baseline settings of our framework. First, we identify high-risk students as those below the 26.25th percentile in the distribution of predicted test scores, which gives a high-risk student share of 0.2625 (rounded to 0.262 in the table). From equation (3), with $B = 1.25N$, the third row of the table shows that $Z = 0.952$. The policy simulation allocates $(1+Z)$ resource units to each student identified as high risk—in the bottom panel of the table, we show the tautological result that students identified as high risk based on a low value of $\hat{S}_i$ each receive 1.952 resource units.

The other rows in the bottom panel of Table 4 show the average resource units accruing to students with other characteristics. For example, students identified as high-risk based on actual test scores (i.e., with S_i below the 26.25th percentile) receive 1.537 resource units, on average. This value is below the value for students identified by $\hat{S}_i$ because the model does not predict test performance perfectly. DC and FRM students receive 1.50 and 1.40 resource units on average, respectively, and the values accruing to ELL, IEP, and URM students are similarly shown. The resources accruing to students with different characteristics derive from the association of these characteristics with low predicted academic performance (i.e., $\hat{S}_i$).

The normalization of resource units in our policy simulation facilitates straightforward comparisons within and across columns in the table. The easiest way to compare allocations across student types is in percentage units relative to the normalized baseline allocation of 1.0, which in a funding system would correspond to a foundational dollar value per pupil. For example, in the baseline scenario in column (1) our framework allocates 1.621 resource units per URM student, on average, or an additional 62.1 percent of the foundational amount received by a low-risk student.

Next, we turn to the comparative analyses in Scenarios 2 and 3. In these scenarios we anchor our framework to the DC and FRM data, respectively, by resetting $\tilde{S}$ to match the share of students identified as high-risk by these designations. That is, 27.3 and 50.3 percent of

Missouri students are directly certified and FRM-eligible, respectively, and we adjust $\tilde{S}$ so the bottom 27.3 and 50.3 percent of students based on $\hat{S}_i$ are identified as high-risk. This allows for comparisons of how the use of our risk measures differs from the alternative measures, holding fixed the fraction of high-risk students identified (and correspondingly, the value of *Z*).

First, the results from Scenario 2 show that using a DC-based definition of risk results in more resources accruing to DC and FRM students, on average, compared to defining risk using $\hat{S}_i$. The finding for DC students is again tautological—when we define risk using DC status, each DC student receives (*1+Z*) resource units by construction. The finding for FRM students follows from the strong overlap between FRM eligibility and DC status. However, the other rows of the table show that targeting resources directly to DC students comes at the cost of lower per-pupil resources for other types of students at risk of low academic performance. First, and unsurprisingly, our framework in Scenario 2 allocates more resources to students with low test scores and low predicted test scores (where the latter reflects the same tautology described above). This empirically confirms that our framework is more effective at targeting students at risk of poor academic performance than a DC-based framework. Our framework also allocates more resources to ELL, IEP, and URM students, and by a substantial margin in all three cases.

A similar set of results unfolds in Scenario 3, which is anchored to FRM status. The magnitudes of the per-student resource allocations in Scenario 3 are smaller across the board compared to Scenario 2 because *Z* is much smaller. This is owing to the fact that many more high-risk students are identified (because there are so many students who are FRM-eligible), which suppresses the per-student allocations under the fixed budget. Still, the general pattern of findings from Scenario 2 holds in Scenario 3. The FRM-based system is, by definition, better at targeting resources to FRM students, and similar or worse at targeting resources to every other category associated with student risk.

Next, we compare columns (1) and (5) across scenarios. This comparison pits our framework under the baseline settings against the FRM-based alternative. The conditions in columns (1) and (5) differ by both (a) the metric used to identify high-risk students and (b) the share of students identified. The comparison shows that our framework allocates more resources per student along every measured dimension of risk except FRM status (including DC status, albeit marginally). For most non-FRM characteristics, our framework leads to substantially more

resources per student, on average. This reflects the broader targeting of resources in our framework based on the full vector of information, **X**, and the fact that we identify fewer high-risk students, which permits a greater per-student allocation via the higher value of *Z*. A summary characterization of this comparison is as follows: our framework is more effective at targeting resources toward high-risk students both as we define them in terms of academic performance and more broadly using most other common categorical definitions.

The comparative scenarios in Table 4 are informative but generic. In Table 5, we show results from a complementary comparison grounded in a real-world policy by grafting the core features of California's high-profile Local Control Funding Formula (LCFF) onto the Missouri data—namely, LCFF's supplemental and concentration grants. This allows us to compare resource allocations based on our framework to what they would look like if LCFF were implemented in Missouri.

California's LCFF allocates additional resources to "targeted disadvantaged pupils" as identified by ELL status, FRM status, and foster youth. Students who belong to any category are counted and students cannot be double-counted based on assignments to multiple categories.[20] We implement a modified version of LCFF that ignores foster youth because we do not have access to data for this designation.[21]

The LCFF allocates resources at the district level and accounts for district-level circumstances by providing additional per-pupil funding to districts with concentrated need (Johnson and Tanner, 2018). Based on 2021 LCFF funding rules, we convert the LCFF formula from a district-level to student-level allocation model to fit within our analytic framework. The student-level version of the LCFF funding formula is as follows:

$$F_i = F_0 + (0.2 * F_0) * D_i + (0.65 * F_0) * \max[D_d - 0.55, 0] \quad (4)$$

In equation (4), F_i is the resource allocation for student *i*, F_0 is the base amount, D_i is an indicator equal to one if the student belongs to a "targeted-disadvantage" category (i.e., ELL or FRM), and D_d is the share of students in a targeted disadvantage category in district *d*. In words,

[20] See here for more information about LCFF (link retrieved 11.01.2021): https://www.cde.ca.gov/fg/aa/lc/lcffoverview.asp; also Johnson and Tanner (2018).
[21] The implications of omitting foster youth should be small because few children in Missouri (1.4 percent in 2020) are in foster care (link retrieved 11.01.2021: https://www.stltoday.com/news/local/state-and-regional/missouri-foster-parents-get-help-from-legislature-but-why-are-more-children-coming-into-state/article_24fab000-d8ed-5ff7-a20d-eaa88a1ae21f.html) and of those that are, many are likely already FRM-eligible.

the LCFF allocates an additional 20 percent of the base funding level for each targeted student, then an extra 65 percent of the base amount to districts for each targeted student in excess of 55 percent of enrollment.[22] Following our analytic structure from above, we normalize F_0 to 1.0.

We apply the pseudo-LCFF in Missouri and assign students values of F_i. To compare the subsequent student allocations to allocations from our framework, we first use the sum of the F_i values across all students to calculate the total pseudo-LCFF budget in Missouri—i.e., the total amount allocated to students under the LCFF rules—which we set as *B* in equation (2). For notational convenience, we write the total budget in units of *N* as above (applying the LCFF rules in Missouri generates a total resource budget of 1.152*N*). Then, using this budget, we implement our policy simulation where we allocate resources following equation (2) and set $\tilde{S}$ at the 26.25th percentile of test scores. This facilitates a fixed-budget comparison between the two allocation models.

Before turning to the results, we note two key features of the pseudo-LCFF. First, it accounts for two categories of disadvantage simultaneously (FRM and ELL), albeit simply. Second, the "concentration" portion of the pseudo-LCFF formula allocates more resources to districts with concentrated need via the third term in equation (4). Our policy structure does not include a directly-analogous concentration component, but the use of the school-level variables in our prediction model is similarly-spirited. That is, to the extent that concentrated student risk is associated with lower test scores conditional on students' individual risk, our model will assign students in high-concentration schools lower values of $\hat{S}_i$. (We could also modify our policy structure to copy the LCFF concentration-grant structure by forcefully allocating more resources to students in schools or districts with high proportions of low-$\hat{S}_i$ students, although we do not pursue this extension here.)

Table 5 shows the results comparing our framework to the pseudo-LCFF in Missouri. Along most dimensions we measure, including the key metrics of test performance and predicted test performance, our framework yields more resources per high-risk student than the pseudo-LCFF. The one exception is FRM students, who are explicitly targeted by LCFF and receive

[22] The way the student-level formula is written in equation (4), each student in a district with $D_d > 0.55$ receives a small positive increment, which is equivalent to identifying the fraction of students above 55 percent and providing the district with the full increment for each of these students.

modestly higher resources under LCFF, on average.[23] Our framework yields higher per-student allocations along most dimensions because it explicitly accounts for them in constructing the predicted test scores. Moreover, it distributes the excess budget (0.152*N) in a more targeted way by focusing on the bottom 26.25 percent of students; in contrast, under the pseudo-LCFF the excess budget is distributed across 51.1 percent of Missouri students (the unduplicated sum of FRM and ELL students).

Tables 4 and 5 focus on student level allocations, but it is difficult to target resources to individual students differentially within a school. The extent to which student-level changes in resource allocations will impact school-level allocations, whether in our framework or in any other framework, depends on the distribution of student characteristics across schools. As a simple example, consider a hypothetical (and unrealistic) setting where students are distributed to schools randomly. In this case, there would be no expected effect of changes to student-level resource allocations on *school-level* resource allocations. This is because each school's student body would be of the same proportions (subject to sampling variance). In the real world, residential sorting implies that changes to student-level allocations will translate at least partly to changes in school-level allocations.

In Appendix C we provide information complementary to Tables 4 and 5 at the school level, in the form of correlations between school-average student characteristics and school-average resource allocations. The correlations reported in the appendix are directionally in line with expectations based on the findings in Tables 4 and 5 in the presence of residential sorting. That is, policy simulations that produce higher *student-level* resources for particular types of students also generally produce higher *school-level* resources for schools serving more of these students. Interestingly, the concentration portion of the pseudo-LCFF formula does not seem to greatly affect the strength of the school-level correlations, although it does put modest upward pressure on the correlation between resources and schools' FRM shares.

The link between our student-level and school-level findings is conditional on the distribution of students across Missouri schools. For our purposes it is sufficient to show that directionally, the correlations in the appendix are as expected based on the student-level findings

[23] ELL students are also explicitly targeted by the pseudo-LCFF, but the effect on ELL students is overwhelmed by other factors. The ELL comparison in Missouri is also not especially useful due to the low ELL share in the state (in contrast to California).

reported in Tables 4 and 5. But more broadly, the question of how student-targeted resource allocation models impact school (and district) resources through the distribution of students across schools (and districts)—whether in our framework or a different framework—merits attention in future research.[24]

6. Framework Flexibility & Augmentation

6.1 Flexibility

An advantage of our framework is that it can handle changes in the underlying variables used to measure student risk, or the information they contain, with greater flexibly and less disruption than systems that rely on categorical assignments. For example, consider a state that measures risk categorically using DC status. If that variable were to suddenly become unavailable (e.g., due to a data sharing problem or other policy change), the entire category would disappear and need to be replaced. This would likely result in a significant disruption to measurement. In contrast, the effect of the same event on our risk measures, $\hat{S}_i$, will be dulled by the larger framework. The reason is that the remaining predictors in the **X**-vector will absorb some of the information loss—i.e., to the extent that these variables are correlated with DC status, their weights in the coefficient vector $\boldsymbol{\beta}_1$ will change to lessen the total impact on the predictions.

This is a clear theoretical benefit of our approach, but does it help in practice? We answer this question in two ways. First, in Table 6 we report correlations of $\hat{S}_i$ as estimated by the specifications shown in Table 2. The correlations are shown in reverse order for the specifications in rows (*l*) to (*a*) of Table 2 in order to emphasize differences among our preferred models that use the richest control-variable sets. Column 1 shows that compared to our primary specification in row (*l*), most alternative specifications of the prediction model yield similar risk values for individual students. For example, only 3 of the 11 correlations reported in the first column are below 0.89, and the minimum value is 0.844 (from row (*a*), which is the sparsest specification). In fact, the correlation reported between models (*a*) and (*l*) of 0.844 is the

[24] IEP students offer an instructive example of the importance of the student distribution because they are more evenly distributed across schools than most other student groups (for whom residential segregation is stronger). Because of their relatively even distribution, as student-level resources for IEP students increase, it can put downward pressure on cross-school differences in resources.

minimum value in the entire correlation matrix. Broadly speaking, Table 6 confirms that the risk metrics for individual students are fairly stable as we change the attributes included in **X**.

Next, in Table 7 we make a concrete comparison between our framework and a categorical system. We assess the implications of a switch from using DC status to using *free meal* (FM) status to identify students from low-income families. Conceptually this is a reasonable substitution as both metrics are purported to identify students from families at 130 percent of the poverty line or below (although in practice FM-eligibility is oversubscribed—see Domina et al., 2018; Fazlul, Koedel, and Parsons, 2021). In the categorical system, we recode students as high risk based on FM status instead of DC status. Within our framework, we make the same data switch in the prediction model. For the DC-data scenario we estimate the model as described in row (*l*) of Table 2 precisely; for the FM-data scenario we estimate the same model but for any DC-based variable or interaction, we use an FM-based variable or interaction in its place. We continue to identify at-risk students in our framework based on predicted achievement—i.e., an at-risk student has $\hat{S}_i < \tilde{S}$, where $\tilde{S}$ is set at the 26.25$^{\text{th}}$ percentile.

The results in Table 7 make clear that the flexibility of our approach is a significant practical benefit. The risk metrics from our framework are much less volatile than the categorical alternative in response to the hypothetical switch from DC to FM data. Specifically, in our framework this data switch results in just 4.4 percent of students switching at-risk status, compared to 17.8 percent of students under the categorical alternative.

The reason for this difference is that the weighting parameters in the prediction model adjust to reflect the informational content of the new variable, in this case FM status, holding the share of at-risk students fixed. We largely identify the same group of at-risk students regardless of whether we use DC or FM data. The change in the categorical designations is much more disruptive because of the large difference in the size of the DC and FM categories and the inherent inflexibility of the categorical approach. We acknowledge that the categorical approach in Table 7 is a straw man in the context of an academic investigation—it is obvious that it must perform worse than the model-based approach—but it is important to recognize this is a fundamentally accurate characterization of current systems and explains much of the consternation brought about by the Community Eligibility Provision, which is pushing states to switch from FRM-based to DC-based categories.

By the same logic, changes to variable definitions that alter their informational content

will also have less of an impact in our framework. For example, suppose future changes to the programs that determine DC status change the level of poverty associated with this variable. The risk designations of students in our framework will not be entirely impervious to such a change, but the impact of the information disruption will be reduced.

We also note one aspect of our measures that makes them more volatile than categorical alternatives: their anchoring to state assessments. If states change their assessments, it will impact students' $\hat{S}_i$ values if the **X**-vector attributes differentially predict performance on the new assessments. We do not explore the sensitivity of our estimates to test changes directly, but make three comments about this issue. First, if a test change occurs and it alters the values of $\hat{S}_i$, it can be argued that the new values of $\hat{S}_i$ reflect an update to the state's educational goals per their decision to adopt the new assessment. Given this, our framework's potential sensitivity to the new test could be viewed positively as it allows for a fluid, policy-aligned adjustment to what it means to be at risk. Second, if a test-induced change to students' $\hat{S}_i$ values is deemed too disruptive, the impact can be dampened by using a multi-year average of the weighting parameters, $\boldsymbol{\beta_1}$, based on the new and old tests. Finally, although there is no direct research on the impact of test changes on prediction models such as ours *per se*, Backes et al. (2018) study the sensitivity of estimates of teacher value added to test changes. Estimates of teacher value added derive from models that are similar in many respects to our prediction models, and in some ways would be expected to be affected even more by test changes. However, Backes et al. (2018) show that estimates of teacher value added are relatively stable when test changes occur, which suggests that the risk measures that emerge from our framework may not be highly sensitive to assessment changes.

6.2 Augmentation

There are some aspects of current systems that our framework does not meaningfully improve upon. The most obvious example is students with severe IEPs, for whom broad categorical designations, or measures of risk from our framework, are insufficient to capture the extent of their needs. This is despite the fact that IEP status is strongly associated with low test performance conditional on taking the test, as demonstrated above by the large average allocations that accrue to IEP students in our framework.

In funding policies, add-ons will be needed for IEP students to augment any general

framework. For accountability policies, which we discuss briefly in an extension below, it may be desirable to exclude these students, or at least those whose disabilities are deemed severe enough to exempt them from testing, as is common practice in many states.

7. Extensions

7.1 Risk Measures that Ignore Race-Ethnicity

The goal of the prediction models summarized by Table 2 is to predict academic performance, and race-ethnicity is a consistently strong predictor. Statistically, the decision of whether to use information on race-ethnicity to improve the estimates of $\hat{S}_i$ is unambiguous: these data should be used. Still, arguments have been made for the omission of data on race-ethnicity from test prediction models such as ours, based on the view that controlling for race-ethnicity sets different expectations for academic performance across racial-ethnic groups.[25] We believe this viewpoint is misguided in our application and that our use of racial-ethnic data in the prediction model acknowledges longstanding gaps in educational outcomes by race-ethnicity, gaps that policies informed by the risk measures from our framework can work to remedy. Still, in this section we briefly consider the implications of the omission of these variables from the prediction model.

We re-estimate our preferred specification in row (*l*) of Table 2 omitting all information about race-ethnicity. This produces estimates of $\hat{S}_i$ that are not directly influenced by this information (although some circumstances that contribute to variance in $\hat{S}_i$ are still correlated with race-ethnicity). Appendix Table D1 shows results from this model in the same format as Table 2. A comparison between the versions of model (*l*) that do and do not include the race-ethnicity variables shows that the predictions from the latter are clearly worse. For example, the R-squared is 0.03 points lower, the MSE is 0.02 points higher, and the classification error rate is 1.5 percentage points higher.

Next, in Appendix Table D2 we use our policy simulation to show how student-level resource allocations are affected if we use the values of $\hat{S}_i$ from the restricted model. The results can be compared to the findings from our baseline scenario in Table 4, which is replicated in the appendix for ease of presentation. Most of the findings are similar regardless of whether we use

[25] Ehlert et al. (2016) provide a deeper discussion on this issue in the context of school accountability systems.

the full or restricted versions of model (*l*), which is as expected given the general robustness of the prediction framework shown by Table 6. However, there is one exception precisely where it is anticipated: using the model that is stripped of all racial-ethnic information results in less resources accruing to URM students.

7.2 Monitoring Achievement Gaps

The empirical application of our framework is in the context of a resource-allocation policy. However, our framework also has features that make it appealing for use in other policies, such as for monitoring achievement gaps within schools. Many states informally monitor within-school achievement gaps and these gaps are incorporated into some states' formal accountability policies (Martin, Sargrad, and Batel, 2016).

Based on their plans submitted to the federal government as part of the Every Student Succeeds Act (ESSA), states currently track achievement gaps in one of two ways. The first is to specify multiple categories of student risk (e.g., FRM, ELL, IEP, URM) and track gaps for each category separately. The second is to combine the categories into one "super subgroup" and track the achievement gap between students who do and do not belong to the super subgroup.

Each approach has strengths and weaknesses. The former follows from the structure of the predecessor to ESSA—No Child Left Behind (NCLB). On the one hand, it is useful because it provides detailed information about achievement gaps along a variety of dimensions. But on the other hand, it can be misleading because of heterogeneity in expected student performance within the categories across schools. For example, if schools A and B both have ELL students, but the ELL students at school A are also at relatively greater risk along other dimensions (e.g., if they come from lower-income families), the ELL-based gap will be higher in school A than in school B due to compositional difference, all else equal.

Another problem with the multi-category approach is that the multiple comparisons can cause information overload.[26] They can also lead to type-I errors because as the number of groups tracked for accountability increases within a school, the likelihood of bad outcomes for some groups by chance increases statistically (Davidson et al., 2015). Policymakers may have trouble drawing accurate inference about schools that track many achievement gaps due to their diversity. The super-subgroup approach is meant to solve these problems by reducing the achievement gap within a school to a single number comparing students who do and do not

[26] See Sutcliffe and Weick (2009) for a general discussion of information overload and its effects.

belong to the super subgroup. However, its limitation is that there are compositional differences in the super subgroup across schools, which exacerbates the problem raised in the preceding paragraph of group heterogeneity in expected student performance.

Our framework allows for the single-comparison simplicity of the super-subgroup approach with the added benefit of minimizing the potential for misleading comparisons due to differences in the composition of super subgroup across schools. The basic idea is to compare schools' predicted achievement gaps between high-risk and low-risk students to their actual gaps. Schools with actual gaps that are smaller than the predicted gaps have less inequity than would be implied by the characteristics of their student bodies, and vice versa for schools with actual gaps that are larger than their predicted gaps. Appendix E provides additional details.

7.3 Uncoarsened Risk Measures

Throughout our empirical application we assign binary risk categories to students based on predicted test performance. This facilitates a straightforward comparison to *status quo* systems and lends policy relevance to our work given the strong cultural norm within the education sector of grouping students categorically, and often in a binary fashion. However, our framework produces more differentiated risk measures in the form of the underlying $\hat{S}_i$ values. This is another dimension of flexibility of our framework.

In the interest of brevity, we do not investigate the potential for using the uncoarsened $\hat{S}_i$ values to enhance policy practice here. A productive avenue of future research would be to consider how using multiple risk categories—e.g., moving from a two-category binary system to a five-category system—could improve resource targeting by facilitating the allocation of additional resources to the highest-risk students. The limiting case would involve using the fully uncoarsened $\hat{S}_i$ values directly in a resource-allocation function to produce even more differentiated allocations.[27]

8. Conclusion

We develop and test a new framework for identifying at-risk students. Our framework is guided by a clear definition of student risk based on predicted academic performance and

[27] Such an exercise may yield useful theoretical insights, although it would be less policy relevant (at least in the near term) given the predominant category-based policy infrastructure in education. In addition to being of less direct use in policy, there are also analytic challenges associated with developing a system based on the fully uncoarsened $\hat{S}_i$ values, some of which we touch on briefly in Appendix B.

modernizes the approach to risk measurement methodologically. The resulting measures are more effective at identifying students at risk of poor academic performance. Our framework is more flexible than *status quo* systems and less sensitive to disruptions caused by changes to the data available for the purpose of risk measurement. The NSLP's Community Eligibility Provision is a recent example of such a disruption. Finally, our framework is designed for use in consequential education policy applications. Although the risk measures it produces are not perfectly non-manipulable, which is the theoretical ideal, the data and estimation procedures outlined in our article aim to minimize their manipulability.

We view our primary contribution as putting forth a principled, methodologically-modernized framework for measuring student risk. We motivate the need for our framework by the lack of a strong conceptual or methodological grounding of current state systems. The historical evolution of these systems is not entirely clear, and we are not aware of any formal documentation of how they came to be. But regardless of the factors that have resulted in current systems, it is difficult to argue they are carefully considered and use available data to measure student risk efficiently. This is a troubling state of affairs given the high stakes associated with accurate risk measurement. Two of the most important types of education policies—funding and accountability policies—depend critically on our ability to identify at-risk students.

We apply and test our framework using the Missouri SLDS in a proof-of-concept exercise. We recognize our decisions about which variables to use as predictors and outcomes, and how to construct some predictor variables (e.g., the panel variables), are subject to reasonable disagreement. But the goal of our paper is not to be prescriptive with regard to the precise details of implementing our framework. In fact, some of our findings refute the notion that there is a clear "right way" to implement the framework that would merit a prescriptive recommendation, which we view as a feature, not a bug (e.g., see Table 6).

Once implemented, our framework is well-suited for continual improvement, which is another advantage over current systems. For example, the set of predictor variables can be augmented in real time as new and higher-quality data become available. It will be important to monitor the potential for measurement disruptions from this kind of augmentation from year-to-year, but the basic diagnostics we present using the Missouri data suggest that the risk measures would not change dramatically in response to most changes to the data. Future iterations of the framework could also incorporate projections of student risk along other dimensions, such as in

terms of attendance, graduation, and college matriculation. These could replace test scores in the framework or, more likely, augment them—for example, each student's total risk score could be a weighted average of risk as assessed for different indicators of academic performance. The framework could even be extended to incorporate emerging measures of student well-being, such as social-emotional measures. Extensions along these lines would require research to assess their costs and benefits, but the flexibility inherent to the framework allows for these kinds of continual improvement efforts. In contrast, the rigidity of existing categorical systems may help to explain why risk measurement in education has not changed for so long.

The impetus for the development of our framework is the inadequacy of current methods for measuring student risk. No new approach, including ours, will perfectly measure risk due to the inherent difficulty of the task. But despite their limitations, new systems can improve upon existing systems, and ultimately increase the efficacy of policies designed to promote educational equity.

References

Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25(1), 95-135.

Austin, W., Figlio, D., Goldhaber, D., Hanushek, E., Kilbride, T., Koedel, C., Lee, J. S., Lou, J., Özek, U., Parsons, E., Rivkin, S., Sass, T., Strunk, K. (2021) Academic mobility in U.S. public schools: Evidence from nearly 3 million students. CALDER Working Paper No. 227-0821-2.

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

Bass, D. N. (2010). Fraud in the lunchroom? *Education Next, 10*(1), 67-71.

Boyd, D., Lankford, H., Loeb, S., and Wyckoff, J. (2013). Measuring test measurement error: A general approach. *Journal of Educational and Behavioral Statistics* 38(6), 629-663.

Chingos, M.M. (2018). A promising alternative to subsidized lunch receipt as a measure of student poverty. Policy report. Washington DC: Brookings Institute.

Data Recognition Corporation. (2019). Missouri assessment program grade-level assessments: English language arts and mathematics grades 3-8 and science grades 3 and 5. Technical report 2019. Maple Grove, MN: Data Recognition Corporation. (retrieved 07.20.2021 at https://dese.mo.gov/college-career-readiness/assessment/assessment-technical-supportmaterials)

Davidson, E., Reback, R., Rockoff, J., and Schwartz, H.L. (2015). Fifty ways to leave a child behind: Idiosyncrasies and discrepancies in states' implementation of NCLB. *Educational Researcher* 44(6), 347-358.

Domina, T., Pharris-Ciurej, N., Penner, A.M., Penner, A.K., Brummet, Q., Porter, S.R., & Sanabria, T. (2018). Is free and reduced-price lunch a valid measure of educational disadvantage? *Educational Researcher* 47(9), 539-555.

Ehlert, M., Koedel, C., Parsons, E., and Podgursky, M. (2016). Selecting growth measures for use in school evaluation systems: Should proportionality matter? *Educational Policy* 30(3), 465-500.

Fazlul, I., Koedel, C., and Parsons, E. (2021). Free and reduced-price meal eligibility does not measure student poverty: Evidence and policy significance. CALDER Working Paper No. 252-0521.

Goldhaber, D., Koedel, C. Özek, U., and Parsons, E. (2022). Using longitudinal student mobility to identify at-risk students. *AERA Open* 8(1).

Goldhaber, D., and Özek, U. (2019). How much should we rely on student achievement as a measure of success? *Educational Researcher* 48(7), 479-83.

Greenberg, E. (2018). New measures of student poverty: Replacing free and reduced-price lunch status based on household forms with direct certification. Education Policy Program policy brief. Washington DC: Urban Institute.

Johnson, R.C., and Tanner, S. (2018). Money and freedom: The impact of California's school finance reform on academic achievement and the composition of district spending. Technical Report. Getting Down to Facts II. Palo Alto, CA: Policy Analysis for California Education.

Koedel, C., and Parsons, E. (2021). The effect of the community eligibility provision on the ability of free and reduced-price meal data to identify disadvantaged students. *Educational Evaluation and Policy Analysis* 43(1), 3-31.

Konstantopoulos, S., and Borman, G.D. (2011). Family background and school effects on student achievement. A multilevel analysis of the Coleman data. *Teachers College Record* 113(1), 97-132.

Li, Y., Scala, J., Gerdeman, D., & Blumenthal, D. (2016). District Guide for Creating Indicators for Early Warning Systems. San Francisco: REL West at WestEd.

Martin, C., Sargrad, S., and Batel, S. (2016). Making the grade: A 50-state analysis of school accountability systems. Policy Report. Washington DC: Center for American Progress.

Michelmore, K., & Dynarski, S. (2017). The gap within the gap: Using longitudinal data to understand income differences in educational outcomes. *AERA Open* 3(1), 1-18.

Public Impact with Education Analytics (2021). Identifying schools achieving great results with highest-need students: Catalyzing action to meet the needs of all students. Chapel Hill, NC: Public Impact. Retrieved 12.06.2021 from https://publicimpact.com/wp-content/uploads/2021/03/Identifying_Schools_Achieving_Great_Results_with_Highest-Need_Students.pdf

Sutcliffe, K. M., & Weick, K. E. (2009). Information Overload Revisited. In *The Oxford Handbook of Organizational Decision Making* (eds. Gerard P. Hodgkinson and William H. Starbuck). Oxford, UK: Oxford University Press.

Table 1. Descriptive Statistics for Missouri Students, 2017.

	Mean	SD
Demographics		
Female	0.49	0.50
American Indian	0.00[a]	0.07
Asian/ Pacific Islander	0.02	0.15
Black	0.16	0.37
Hispanic	0.06	0.24
White	0.72	0.45
Multi-race	0.03	0.18
English Language Learner	0.04	0.20
Individualized Education Program	0.13	0.34
Poverty Measures		
Directly Certified	0.27	0.45
Free and Reduced-Price Lunch Eligible	0.50	0.50
Free-Lunch Eligible	*0.44*	*0.50*
Reduced-Price Lunch Eligible	*0.06*	*0.24*
Mobility Measures		
Number of Districts Attended	1.04	0.22
Number of Schools Attended	1.05	0.24
Test Scores (Standardized)		
Average Math and English Language Arts	0.01	0.92[b]
N (students)	698,726	

Notes: This table shows the summary statistics for students in Missouri in the 2016-2017, restricted to students in schools with at least 25 students enrolled. Test scores are from a reduced sample of 387,317 students in grades 3-8 with math and communication arts tests.

[a]0.4 percent of Missouri students are American Indian

[b] The standard deviations of the standardized math and English Language Arts tests in the analytic sample are 0.99 separately; the standard deviation of students' averaged standardized scores is lower.

Table 2. Statistical Output from Various Test Prediction Models.

	R-squared from predictive linear regression	MSE	Classification error rate percentage (i.e., predicted status ≠ actual status)		
	(1)	(2)	(3)	(4)	(5)
Predicting students' contemporary test scores using:			All	False positive	False negative
(a) Individual contemporary variables	0.213	0.67	23.56	8.92	14.64
(b) Individual contemporary variables with FRM	0.266	0.62	24.12	12.55	11.57
(c) Individual contemporary variables with DC	0.248	0.64	23.72	11.10	12.62
(d) Individual contemporary variables with DC and two-way interactions	0.251	0.63	24.32	12.79	11.53
(e) All individual variables in (a), plus corresponding panel variables	0.221	0.66	23.90	10.86	13.04
(f) All individual variables in (b), plus corresponding panel variables	0.277	0.61	23.93	12.18	11.75
(g) All individual variables in (c), plus corresponding panel variables	0.259	0.63	24.12	12.40	11.72
(h) All individual variables and two-way interactions in (d), plus corresponding panel variables and two-way panel interactions	0.263	0.62	24.09	12.53	11.56
(i) All individual and panel variables in (e), plus corresponding school-level aggregates	0.250	0.63	24.21	12.31	11.90
(j) All individual and panel variables in (f), plus corresponding school-level aggregates	0.290	0.60	23.81	12.20	11.61
(k) All individual and panel variables in (g), plus corresponding school-level aggregates	0.282	0.61	24.09	12.81	11.28
(l) All individual and panel variables and two-way interactions in (g), plus corresponding school-level aggregates and two-way school level interactions	0.290	0.60	23.81	12.59	11.22
N (Test Takers in Grades 3-8)	387,317				
N (Schools)	1,749				

Notes: Rows (a) – (d) include individual contemporary variables for students. Row (a) includes information about mobility, EL status, IEP status, sex, and race-ethnicity indicators. Row (b) adds FRM status to the variable list in row (a), and row (c) replaces FRM status with DC status. Row (d) includes all the variables in row (c) and adds all possible two -way interactions of these variables. Rows (e) to (h) include individual level panel variables corresponding to those in rows (a) – (d). Row (e) adds three-year averages of school and district mobility, share of years spent in a Missouri public school in the last three years as well as separate variables indicating the share of the last three years spent as an EL and IEP student. Model (f) adds the share of years as an FRM student, model (g) replaces that with DC status panel variable, and model (h) adds two-way interactions for all panel variables used in model (g), along with previous interactions of the individual variables. Finally, models (i) – (l) add school level aggregate variables to models (e) – (h) in the same fashion. The R-squared values indicate the share of the variance in the outcome—in this case, the student's year-t standardized test score averaged over math and communication arts that can be explained by the variables in each row. The binary classification error rates are calculated as the fraction of students whose predicted binary proficiency classification differs from their actual classification based on their observed test scores.

Table 3. Means and Standard Deviations of $\hat{S}_i$ Overall, and Within Traditional Categories of Disadvantage.

	All students	DC		FRM		ELL		URM		IEP	
		DC	Not DC	FRM	Not FRM	ELL	Not ELL	URM	Not URM	IEP	Not IEP
Full specification (from row (*l*) of Table 2)											
Average $\hat{S}_i$	0.03	-0.39	0.19	-0.23	0.29	-0.39	0.05	-0.43	0.16	-0.78	0.16
Standard deviation of $\hat{S}_i$ (with shrinkage)	0.50	0.44	0.42	0.49	0.35	0.44	0.49	0.49	0.42	0.38	0.39
Share of students in this category	1.0	0.27	0.73	0.50	0.50	0.04	0.96	0.22	0.78	0.13	0.87

Notes: The full specification from which we obtain $\hat{S}_i$ is as shown in row (*l*) of Table 2.

Table 4. Resource Allocation Policy Simulations, Results Part I: Average per-Student Allocations.

	Baseline Scenario: $\tilde{S}$ set at basic/below basic achievement percentile	Scenario 2: $\tilde{S}$ set so the high-risk student share matches the DC share		Scenario 3: $\tilde{S}$ set so the high-risk student share matches the FRM share	
	Use $\hat{S}_i$ to define high risk	Use $\hat{S}_i$ to define high risk	Use DC to define high risk	Use $\hat{S}_i$ to define high risk	Use FRM to define high risk
N(H) Share	0.262	0.273	0.273	0.503	0.503
N(L) Share	0.738	0.727	0.727	0.497	0.497
Z	0.952	0.916	0.916	0.497	0.497
B	1.25*N	1.25*N	1.25*N	1.25*N	1.25*N
Average resource units per student, by type, where a value of 1.0 represents the normalized resource allocation to low-risk students:					
Actual Test Score (S_i) below 26.25th percentile	1.537	1.530	1.445	1.403	1.379
Predicted test score ($\hat{S}_i$) below 26.25th percentile	1+Z=1.952	1+Z=1.916	1.500	1+Z= 1.497	1.400
DC	1.500	1.500	1+Z=1.916	1.482	1.490
FRM	1.400	1.400	1.489	1.391	1+Z= 1.497
ELL	1.636	1.631	1.335	1.456	1.405
IEP	1.910	1.880	1.337	1.494	1.316
URM	1.621	1.618	1.432	1.455	1.404
N	698,726	698,726	698,726	698,726	698,726

Notes: Using different values of *B*, subject to the constraint $B > N$, does not affect the findings directionally, although it does increase the per-pupil dollar gaps for all student categories relative to 1.0.

Table 5. Resource Allocation Policy Simulations, Results Part II: Average per-Student Allocations Under our Framework versus Pseudo-LCFF, Holding the Budget Fixed Based on the Projected LCFF Amount.

	Our Framework	Pseudo-LCFF
N(H) Share	0.262	0.511
N(L) Share	0.738	0.489
Z	0.570	N/A
B	1.152*N	1.152*N
Average resource units per student, by type, where a value of 1.0 represents the normalized resource allocation to low-risk students:		
Actual Test Score (S_i) below 26.25th percentile	1.322	1.235
Predicted test score ($\hat{S}_i$) below 26.25th percentile	1+Z=1.570	1.271
DC	1.299	1.287
FRM	1.239	1.287
ELL	1.381	1.305
IEP	1.545	1.182
URM	1.372	1.295
N	698,726	698,726

Notes: *B* is determined based on the budget implied by the pseudo-LCFF, which we implement as described in the text. We convert the budget into units of *N* to facilitate comparability with other portions of our analysis. The high-risk group under pseudo-LCFF is as defined by that policy: the sum of ELL and FRM (unduplicated).

Table 6. Correlations of $\hat{S}_i$ in the Full Sample When $\hat{S}_i$ is Estimated using Different Variables in the *X*-vector.

	(l)	(k)	(j)	(i)	(h)	(g)	(f)	(e)	(d)	(c)	(b)	(a)
(l)	1.0	--	--	--	--	--	--	--	--	--	--	--
(k)	0.957	1.0	--	--	--	--	--	--	--	--	--	--
(j)	0.917	0.952	1.0	--	--	--	--	--	--	--	--	--
(i)	0.894	0.934	0.920	1.0	--	--	--	--	--	--	--	--
(h)	0.932	0.953	0.908	0.883	1.0	--	--	--	--	--	--	--
(g)	0.923	0.959	0.908	0.887	0.991	1.0	--	--	--	--	--	--
(f)	0.891	0.918	0.976	0.872	0.929	0.929	1.0	--	--	--	--	--
(e)	0.859	0.890	0.873	0.934	0.921	0.927	0.894	1.0	--	--	--	--
(d)	0.910	0.933	0.891	0.877	0.979	0.973	0.912	0.919	1.0	--	--	--
(c)	0.903	0.939	0.890	0.880	0.971	0.979	0.911	0.924	0.992	1.0	--	--
(b)	0.879	0.906	0.956	0.866	0.919	0.920	0.980	0.891	0.923	0.923	1.0	--
(a)	0.844	0.874	0.858	0.917	0.907	0.912	0.879	0.984	0.925	0.931	0.896	1.0

Notes: The row and column headers reference the rows of Table 2 that define the variable list used to estimate $\hat{S}_i$. We use our baseline jackknifing scenario that jackknifes the data into four equal-sized groups of districts to produce these correlations.

Table 7. Changes in Student Categorizations as At Risk in the Hypothetical Condition where DC Data Become Unavailable and the State Switches to Using Free-Meal (FM) Status in its Place.

	Our Framework: $\hat{S}_i$ is predicted first with DC data using the model shown in row (*l*) of Table 2, then using the same model but with FM data in place of DC data; high-risk status in both scenarios is assigned if $\hat{S}_i < \tilde{S}$	Categorical System: At-risk status is initially assigned categorically based on DC status, then by FM status
Share of high risk students using DC	0.262	0.273
Share of high risk students using FM	0.262	0.441
Share of students who have a change in risk status (high to low, or low to high) due to the data change	0.044	0.178

Notes: The first scenario within our framework corresponds to row (*l*) of Table 2; the second scenario is identical except we replace any DC-based information with FM-based information in the prediction model.

Appendices

Appendix A: Supplementary Tables

Appendix Table 1. Correlations of $\hat{S}_i$ in the Full Sample when the Predictive Regression is Estimated using Test Data from Grades 3-8, Grades 3-5 Only, and Grades 6-8 Only.

	$\hat{S}_i$ estimated using data from test takers in grades 3-8	$\hat{S}_i$ estimated using data from test takers in grades 3-5	$\hat{S}_i$ estimated using data from test takers in grades 6-8
$\hat{S}_i$ estimated using data from test takers in grades 3-8	1.0	--	--
$\hat{S}_i$ estimated using data from test takers in grades 3-5	0.977	1.0	--
$\hat{S}_i$ estimated using data from test takers in grades 6-8	0.974	0.930	1.0

Notes: The values of $\hat{S}_i$ are from the primary specification described by row (*l*) in Table 2. We use our baseline jackknifing scenario that jackknifes the data into four equal-sized groups of districts to produce these correlations.

Appendix Table A2. Correlations of $\hat{S}_i$ in the Full Sample Under Different Jackknifing Scenarios.

	"Leave-out-one-quarter" jackknife (baseline)	"Leave-out-one-third" jackknife	"Leave-out-one-fifth" jackknife	"Leave-out-one-tenth" jackknife	"Leave-out-one-district" (pure) jackknife
"Leave-out-one-quarter" jackknife (baseline)	1.0	--	--	--	--
"Leave-out-one-third" jackknife	0.988	1.0	--	--	--
"Leave-out-one-fifth" jackknife	0.987	0.987	1.0	--	--
"Leave-out-one-tenth" jackknife	0.995	0.988	0.990	1.0	--
"Leave-out-one-district" (pure) jackknife	0.983	0.976	0.987	0.984	1.0

Notes: The values of $\hat{S}_i$ are from the primary specification described by row (*l*) in Table 2.

Appendix B: Supplementary Information about the Test Prediction Models

In this appendix we provide additional details about the prediction models beyond what is shown in Tables 2 and 3. We do not report values for the individual coefficients in the prediction models because the multivariate regression framework makes it difficult to gain inference from them, especially in our richer (and preferred) specifications that include overlapping information (e.g., contemporary and panel measures of the same concepts, interactions of variables, etc.).[28] What we do show in Figure B1 is the distributions of predicted scores, $\hat{S}_i$, for the different specifications shown in Table 2 in the main text.

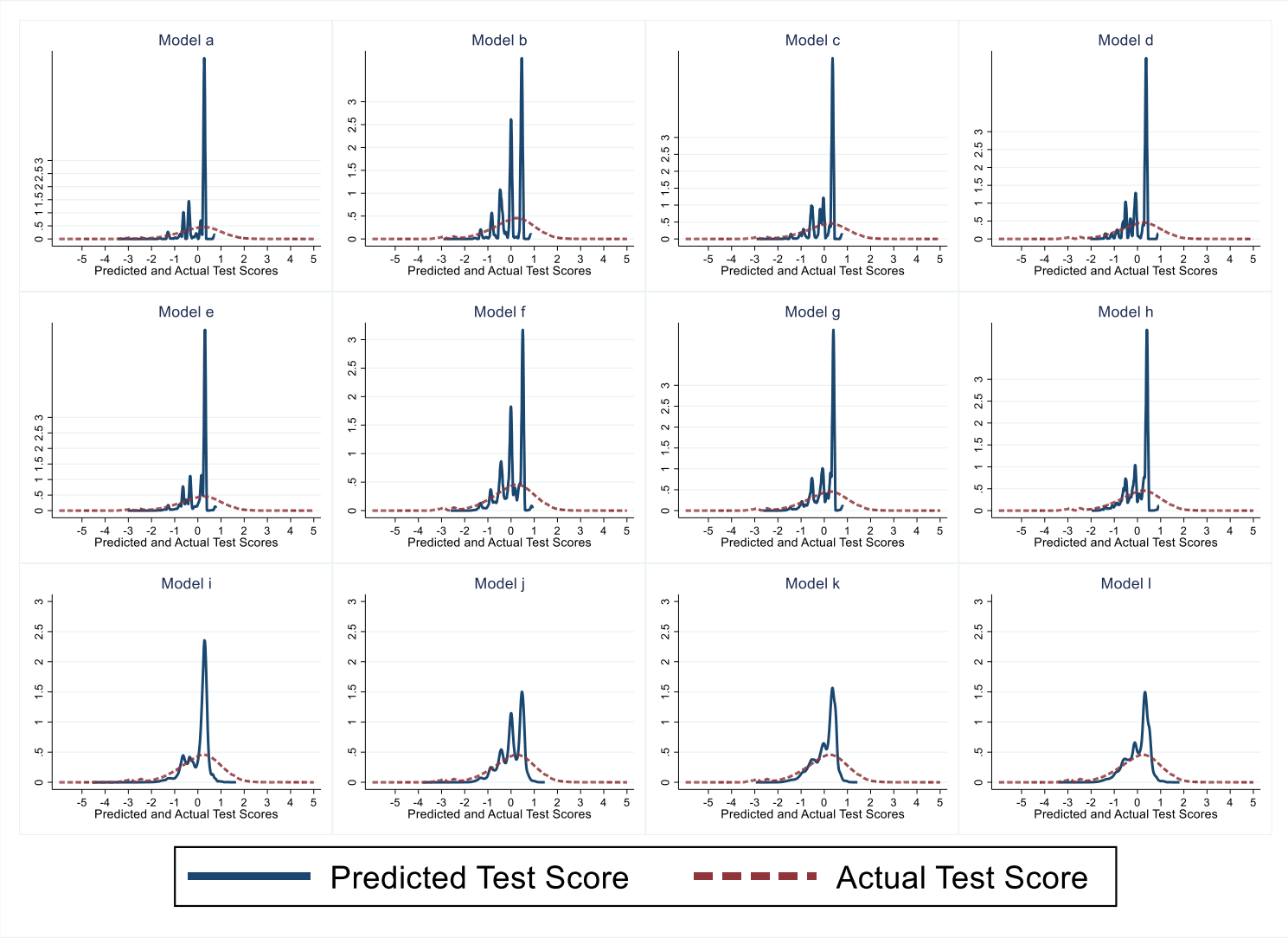
There are two reasons we show the distributions. First, they highlight the "lumpiness" in the distributions of $\hat{S}_i$, especially for the sparser versions of the prediction model. The lumpiness is not surprising because all of the student-level control variables in the models are binary or categorical indicators. When we add the panel versions of the variables to the prediction model it facilitates greater dispersion of the predicted values, and even more so when we add the school-average variables. This explains why the distributions of $\hat{S}_i$ are less lumpy going down the rows of graphs in Figure B1. Still, none of the distributions of $\hat{S}_i$ in the figure are smooth, which reflects the nature of the underlying data.

This is a limitation in the sense that it would be beneficial to have better-differentiated, consequential predictors of student test scores available in state data (i.e., continuous or near-continuous predictor variables of consequence). In the absence of such variables, the distributions of $\hat{S}_i$ are necessarily lumpy. That said, and following on the discussion in the text, it is still true that the degree of differentiation in $\hat{S}_i$ is far greater than the differentiation currently facilitated by states' categorical systems for identifying at-risk students. This is because the model allows students with different combinations of categorical assignments to have different values of $\hat{S}_i$. If useful predictors of student test scores become available in the future that are continuous or near-continuous, they could be incorporated into the framework in a straightforward manner to smooth the predictions further.

[28] Said another way, the "all else equal" interpretation typically ascribed to regression coefficients is not sensible in our models. However, a broad sense of how our predictions associate with key student characteristics is provided in Table 3 in the main text.

The second reason we show the predictions is to make clear that they do a poor job of differentiating students in the upper end of the distribution. This again reflects a feature of the underlying data: namely, that the data available in state systems are insufficient to differentiate high-achieving students. From the perspective of a generic predictive-modeling exercise this is a serious limitation, but for our application it is not because we do not need to differentiate students in the upper end of the distribution to inform policies targeted toward high-risk students. The distributions from the richer specifications in particular show that the prediction model works well in lower tail of the distribution (including our preferred specification, model (*l*)), although this issue will make some potential expansions of our framework problematic—namely, expansions that greatly increase the threshold value for identifying at-risk students, $\tilde{S}$. More broadly, the poor distributional alignment in the upper tail between actual and predicted scores highlights a blind spot in state longitudinal data systems with respect to collecting data that permits the identification of high achievers.

Appendix Figure B1. Predicted Test-Score Distributions Using the Models Shown in Table 2 in the Main Text.



Notes: The graph labels indicate the row in Table 2 to which the model corresponds. The actual test score distribution is the same in each graph

Appendix C: School-Level Allocations

Appendix Tables C1 and C2 provide information complementary to Tables 4 and 5 in the main text, but at the school level. This allows us to assess the extent to which the resource differences across students shown in Tables 4 and 5 translate to cross-school differences, in acknowledgment of the difficulty of targeting resources to individual students differentially within a school.

We start with Appendix Table C1, which replicates the scenarios from the initial policy simulations shown in Table 4. Reflecting this, the first four rows of Table 4 and Appendix Table C1 are identical. The bottom rows differ in that they report correlations between school-average variables and school-average student allocations. Larger correlations, positive or negative, indicate resource allocations that are targeted more or less toward schools that serve students with the characteristics indicated by the rows. The findings in Appendix Table C1 are in the expected direction following on Table 4, although the magnitudes of the correlations vary depending on how students are distributed across schools.

We highlight two key takeaways from Appendix Table C1. First, the tautological aspects of the allocations from Table 4 remain: our framework is better at targeting resources to schools with lower average predicted test scores, and the DC- and FRM-based systems are better at targeting resources to schools with more DC and FRM students, respectively. Second, and also following from Table 4, our framework is more effective at targeting resources to schools with more at-risk students as defined by the non-test-score and non-poverty categories (ELL, IEP, and URM).

Appendix Table C2 reports the same school-level correlations under the pseudo-LCFF. Again, the general insights from the student-level resource allocations in Table 5 are reflected in the school-level correlations. The concentration portion of the pseudo-LCFF formula does not seem to greatly affect the correlations—as evidenced by the substantive similarity of the student-level and school-level results—although it does appear to put modest upward pressure on the correlation between resources and the FRM share.

A formal decomposition of the differences between Tables 4 and 5 and Appendix Tables C1 and C2 is possible and would be useful to quantify how the distribution of student characteristics across schools influences the link between the student-level and school-level

resource allocations. Appendix Table C3 provides the school-level means and variances of the focal student characteristics, which would be key inputs for such a decomposition.[29]

[29] We save this extension for future research because it requires theoretical background and would be more useful if it were conducted so that the results could be generalized broadly (e.g., to settings in different states with different student distributions). We view it as beyond the scope of our current article.

Appendix Table C1. Resource Allocation Policy Simulations, Results Part I.A: Correlations between School-Level Allocations and School Characteristics.

	Baseline Scenario: $\tilde{S}$ set at basic/below basic achievement percentile	Scenario 2: $\tilde{S}$ set so the high-risk student share matches the DC share		Scenario 3: $\tilde{S}$ set so the high-risk student share matches the FRM share	
	Use $\hat{S}_i$ to define high risk	Use $\hat{S}_i$ to define high risk	Use DC to define high risk	Use $\hat{S}_i$ to define high risk	Use FRM to define high risk
N(H) Share	0.262	0.273	0.273	0.503	0.503
N(L) Share	0.738	0.727	0.727	0.497	0.497
Z	0.952	0.916	0.916	0.497	0.497
B	1.25*N	1.25*N	1.25*N	1.25*N	1.25*N
Correlations between average resources and school need as defined by:					
Average test score	-0.697	-0.695	-0.694	-0.659	-0.640
Average predicted test score	-0.829	-0.827	-0.678	-0.771	-0.617
DC share	0.786	0.793	0.994	0.842	0.864
FRM share	0.673	0.681	0.865	0.797	0.998
ELL share	0.227	0.229	0.143	0.200	0.200
IEP share	0.225	0.221	0.161	0.187	0.067
URM share	0.818	0.817	0.630	0.632	0.556
N (schools)	2,101	2,101	2,101	2,101	2,101

Notes: Using different values of *B*, subject to the constraint $B > N$, does not affect the findings directionally, although it does affect the strength of the correlations. The school-level sample size in this table is larger than in Table 2 because Table 2 uses the test-taking sample in grades 3-8 only.

Appendix Table C2. Resource Allocation Policy Simulations, Results Part II.A: Correlations between School-Level Allocations and School Characteristics Under our Framework versus Pseudo-LCFF, Holding the Budget Fixed Based on the Projected LCFF Amount.

	Our Framework	Pseudo-LCFF
N(H) Share	0.262	0.511
N(L) Share	0.738	0.489
Z	0.570	N/A
B	1.152**N*	1.152**N*
Correlations between average resources and school need as defined by:		
Average test score	-0.697	-0.605
Average predicted test score	-0.829	-0.596
DC share	0.786	0.786
FRM share	0.673	0.907
ELL share	0.227	0.223
IEP share	0.225	0.024
URM share	0.818	0.658
N (schools)	2,101	2,101

Notes: *B* is determined based on the budget implied by the pseudo-LCFF, which we implement as described in the text. We convert the budget into units of *N* to facilitate comparability with other portions of our analysis. The high-risk group under pseudo-LCFF is as defined by that policy: the sum of ELL and FRM (unduplicated).

Appendix Table C3. School-Level Average Shares of Students, and their Variances, for Various Student Characteristics.

Measures of disadvantage	Mean of school level share	Variance of school level share
DC	0.323	0.032
FRM	0.577	0.066
ELL	0.040	0.008
IEP	0.152	0.015
URM	0.212	0.083
Low S_i (below 26.25th percentile)	0.283	0.030
Low $\hat{S}_i$ (below 26.25th percentile)	0.304	0.067
N (School)	2,101	

Note: The cross-school mean and variance of observed scores (S_i) are calculated using the subset of schools (1,749) with at least one student who took both the Math and English Language Arts test. Note the school means of the shares of students with low scores based on S_i and $\hat{S}_i$ are not equal to 26.25 because the means in this table are school-weighted, not student-weighted.

Appendix D: Omitting Information about Race-Ethnicity from the Predictions

In this appendix, we briefly report on our findings if we omit information about student race-ethnicity entirely from the prediction model. We do not believe there is a strong rationale for omitting information about race-ethnicity from the model. Nonetheless, we provide this analysis for completeness and in recognition of the fact that some stakeholders may prefer to specify the model in this way, or at least wish to understand the implications.

The results from our analysis omitting all racial-ethnic information from the prediction model are provided in Appendix Tables D1 and D2. Table D1 corresponds to Table 2 in the main text, and Table D2 corresponds to Table 4 in the main text. In very brief summary, Table D1 shows that the prediction model performs worse if we omit racial-ethnic information. This is readily apparent in the output from the predictive regression. The R-squared is lower, the MSE is higher, and the classification error rate is higher. Table D2 shows that for the most part, the average student allocations in the policy simulation are not meaningfully affected by omitting racial-ethnic information from the prediction model (compared to column (1) of Table 4 in the main text). This result follows from Table 6, which shows more broadly that using different predictors, and combinations of predictors, does not have large effects of students' risk-status rankings within our framework. The one exception is with regard to URM status—URM students have much lower average allocations in Table D2 compared to Table 4. This result reflects the fact that if we do not allow for racial-ethnic differences in student performance in the model, it does not recognize race-ethnicity as an independent indicator of risk status; unsurprisingly, this corresponds to fewer URM students being identified as at-risk.

Appendix Table D1. Statistical Output from Primary Test Prediction Model, Omitting All Race-Ethnicity Information.

	R-squared from predictive linear regression	MSE	Classification error rate percentage (i.e., predicted status ≠ actual status)		
	(1)	(2)	(3)	(4)	(5)
Predicting students' contemporary test scores using:			All	False positive	False negative
(l') All variables included in row (*l*) of Table 2 in the main text, except any variables and interactions involving race-ethnicity	0.260	0.62	25.31	14.09	11.23
N (Test Takers in Grades 3-8)	387,317				
N (Schools)	1,749				

Appendix Table D2. Comparison of Primary Policy-Simulation Findings from Table 4 Using Test Prediction Models With and Without Race-Ethnicity Information.

	Baseline Scenario: $\tilde{S}$ set at basic/below basic achievement percentile (Repeated from Table 4)	$\tilde{S}$ set at basic/below basic achievement percentile (Test prediction model does not include any race-ethnicity information)
	Use $\hat{S}_i$ to define high risk	Use $\hat{S}_i$ to define high risk
N(H) Share	0.262	0.262
N(L) Share	0.738	0.738
Z	0.952	0.952
B	1.25*N	1.25*N
Average resource units per student, by type, where a value of 1.0 represents the normalized resource allocation to low-risk students:		
Actual Test Score (S_i) below 26.25th percentile	1.537	1.544
Predicted test score ($\hat{S}_i$) below 26.25th percentile	1+Z=1.952	1+Z=1.952
DC	1.500	1.558
FRM	1.400	1.404
ELL	1.636	1.606
IEP	1.910	1.924
URM	1.621	1.466
N	698,726	698,726

Notes: Using different values of *B*, subject to the constraint $B > N$, does not affect the findings directionally, although it does increase the per-pupil dollar gaps for all student categories relative to 1.0.

Appendix E: Details About Using our Framework for Accountability

In this appendix we provide details about how our framework can be used to improve the monitoring of achievement gaps within schools. Following on the resource-allocation application in the main text, we identify all students with $\hat{S}_i \geq \tilde{S}$ as low risk and all students with $\hat{S}_i < \tilde{S}$ as high risk. We continue with the 26.25th percentile in mind as the threshold for $\tilde{S}$, although this choice is not substantively important in what follows.

Consider the following representation of the observed achievement gap in school *k* between low-risk and high-risk students:

$$\frac{1}{N_{L,k}}\sum_{i=1}^{N_{L,k}} S_{ik} - \frac{1}{N_{H,k}}\sum_{i=1}^{N_{H,k}} S_{ik} \qquad \text{(E1)}$$

In equation (E1), the subscript *k* is added to each student's score, S_{ik}, to denote the school assignment. Next consider the predicted achievement gap based on our framework, where the only change is that we replace students' actual scores, S_{ik}, with their predicted scores, $\hat{S}_{ik}$:

$$\frac{1}{N_{L,k}}\sum_{i=1}^{N_{L,k}} \hat{S}_{ik} - \frac{1}{N_{H,k}}\sum_{i=1}^{N_{H,k}} \hat{S}_{ik} \qquad \text{(E2)}$$

The observed and predicted achievement gaps in equations (E1) and (E2) can be used to determine how the actual achievement gap at school *k* compares to the predicted gap based on the **X**-vector attributes of students who attend school *k*. A useful metric for school *k* can be expressed as the difference between equations (E1) and (E2):

$$\{\frac{1}{N_{L,k}}\sum_{i=1}^{N_{L,k}} S_{ik} - \frac{1}{N_{H,k}}\sum_{i=1}^{N_{H,k}} S_{ik}\} - \{\frac{1}{N_{L,k}}\sum_{i=1}^{N_{L,k}} \hat{S}_{ik} - \frac{1}{N_{H,k}}\sum_{i=1}^{N_{H,k}} \hat{S}_{ik}\} \qquad \text{(E3)}$$

Momentarily suppressing discussion of one technical caveat, equation (E3) has a straightforward interpretation. If the value is positive, the actual achievement gap between low-risk and high-risk students at school *k* exceeds the predicted gap based on the attributes of low-risk and high-risk students; and vice-versa if equation (E3) is negative. Said another way, schools with negative values of equation (E3) have achievement gaps that are smaller than what would be expected based on their compositions of low-risk and high-risk students.

Equation (E3) provides a single, summary indication of how the achievement gap in school *k* compares to what is expected. States can quickly identify schools that have narrower achievement gaps than expected, and larger gaps than expected, based on this single number.

The potential for equation (E3) to be misleading about the school's gap is much less than in the simple systems states currently use. This is because the composition of high-risk and low-risk students along many dimensions is accounted for by the rich specification from which the $\hat{S}_{ik}$ values are estimated.

The one technical caveat to this simple interpretation of equation (E3) is that the fitted values in equation (E2)—i.e., the $\hat{S}_{ik}$ values—are implicitly shrunken through the predictive regression. As noted in the main text, shrinkage is inherent to the prediction process and in the resource-allocation policy simulations, we addressed this issue by using percentiles to set $\tilde{S}$. Due to the shrinkage, the average gap between the test score predictions in equation (E2) will be attenuated relative to the gap in observed scores in equation (E1), resulting in disproportionately positive values for the difference in equation (E3).

Fortunately, as in the allocation-policy context, there are straightforward technical solutions to the shrinkage problem. One solution, following from our preceding analysis, is to estimate equation (E3) using percentiles rather than actual and predicted scores. The interpretation of equation (E3) would be as follows: for each school, it would indicate the difference in the actual versus predicted percentile gap between high-risk and low-risk students. If equation (E3) is estimated in percentiles, the simple interpretation of positive and negative values would hold from above.

However, it may be undesirable from a presentational standpoint for states to report achievement gaps in percentiles. If states wish to report the difference in equation (E3) in test-based units and not percentiles, a mathematically-equivalent solution is to inflate the variance of $\hat{S}_i$ to match the variance of S_i by multiplying the $\hat{S}_i$ values by a constant.[30] This inflation should occur after the predictions are made using equation (1) in the main text, but before constructing the average predicted values in equation (E2). Using the variance-inflated $\hat{S}_i$ values, equation (E3) can be interpreted in test-based units, and the same inference can be drawn for positive and negative values as described above.

[30] Specifically, if each value of $\hat{S}_i$ is multiplied by the ratio of standard deviations of S_i and $\hat{S}_i$, it will inflate the variance so the variance of the modified $\hat{S}_i$ values matches the variance of S_i. This will preserve students' rankings in the distribution of fitted values and allow for appropriate interpretation of equation (E3).