NATIONAL
CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

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

AIR

Using Predicted Academic Performance to Identify At-Risk Students in Public Schools

Ishtiaque Fazlul
Cory Koedel
Eric Parsons

WORKING PAPER No. 261-0922 • September 2022

Using Predicted Academic Performance to Identify At-Risk Students in Public Schools

Ishtiaque Fazlul
Cory Koedel
Eric Parsons

September 2022

Measures of student disadvantage—or risk—are critical components of equity-focused education policies. However, the risk measures used in contemporary policies have significant limitations, and despite continued advances in data infrastructure and analytic capacity, there has been little innovation in these measures for decades. We develop a new measure of student risk for use in education policies, which we call Predicted Academic Performance (PAP). PAP is a flexible, data-rich indicator that identifies students at risk of poor academic outcomes. It blends concepts from emerging "early warning" systems with principles of incentive design to balance the competing priorities of accurate risk measurement and suitability for policy use. PAP is more effective than common alternatives at identifying students who are at risk of poor academic outcomes and can be used to target resources toward these students—and students who belong to several other associated risk categories—more efficiently.

This paper was previously circulated as 'A New Framework for Identifying At-Risk Students in Public Schools'.

Affiliations and Acknowledgement
Fazlul is in the Department of Economics, Finance, and Quantitative Analysis at Kennesaw State University, Koedel is in the department of economics and Truman School of Government and Public Affairs at the University of Missouri, and Parsons is in the Department of Economics at the University of Missouri. 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 data infrastructure investment, 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" (we use these terms interchangeably throughout this article). Today, as has been the case for decades, states ubiquitously rely on blunt categorical indicators associated with disadvantage to identify these students. Examples of common indicators include free and reduced-price meal (FRM) enrollment, direct certification (DC), English language learner (ELL) status, individualized education program (IEP) status, and underrepresented minority (URM) status, among others.

The categorical approach to identifying at-risk students is limited in a number of ways. To illustrate with an example, consider California's Local Control Funding Formula (LCFF), which identifies at-risk students categorically based on whether they are FRM-enrolled or an ELL. If FRM enrollment is a risk indicator and ELL status is a risk indicator, what about a student who is both FRM-enrolled and an ELL? Or a student who is consistently FRM-enrolled versus one who is enrolled for just a single year? California's LCFF—like other categorical systems of which we are aware—does not allow for this type of differentiation to impact students' risk designations, despite its intuitive appeal and clear evidence from research that these differences matter (e.g., see Goldhaber et al., 2022; Michelmore and Dynarski, 2017). This

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

example illustrates two central problems with existing risk measurement systems. First, these systems are not precise about what it means for students to be "at risk"—i.e., there is not a conceptual framework guiding the dimension(s) of risk being measured. Second, they do not leverage the rich information available in state data systems to measure risk as accurately and robustly as possible.

With the limitations of existing measurement practices as motivating context, we contribute to the literature by developing a new framework for measuring student risk in public schools, which we call "Predicted Academic Performance," or PAP. PAP is a singular indicator that draws on the many data elements available in state data systems to measure student risk, defined precisely as "risk of poor academic performance." It blends concepts from emerging "early warning" systems with principles of incentive design to balance the competing priorities of accurate risk measurement and suitability for policy use.

In addition to developing a general measurement framework for PAP, we conduct an empirical proof-of-concept exercise using the Missouri SLDS to understand its potential value. We show that PAP is more effective than *status quo* measures at identifying students who are at risk of poor academic performance, which is by design. Furthermore, in a policy context, we show that PAP can be used to better target resources toward low-performing students and their schools, or relatedly (and with lower stakes), as a diagnostic tool to help policymakers better understand how resources are distributed to students at the greatest risk of poor academic outcomes. We further show that PAP can 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 LCFF.

We show that PAP is a useful measure of student risk and has a number of desirable properties, but it is not a panacea and has limitations that we elaborate on over the course of this article. While we are not claiming to be able to dislodge the entrenched reliance on simple categorical risk indicators in education policies with a single article, we hope that we can propel

research forward on a more promising and modernized path toward the accurate and useful measurement of student risk. Improvements to risk measurement can bear fruit in the form of more efficacious policies designed to narrow achievement gaps and promote better academic outcomes among disadvantaged student populations. Risk measurement is more than a question of measurement—it is a question of policy.

2. Why We Need New Ways to Measure Risk

Indicators of student risk are among the most consequential measures in education policy (along with measures of student performance). These indicators determine how billions of dollars of state funding are allocated to school districts each year through progressive funding formulas in most states.[2] And in conjunction with measures of academic performance, they are the primary tools we use to understand learning gaps and implement policies to mitigate these gaps.

But despite the critical role played by risk indicators in education policies, we know little about how effectively they measure student risk. Consider FRM enrollment, for instance, which is the predominant indicator of "low-income status" in education policies.[3] There has long been a commonsense understanding that FRM enrollment is an imperfect proxy for family income (e.g., see Bass, 2010; Harwell and LeBeau, 2010). In addition, more recently there is definitive evidence of the gross inaccuracy of FRM-based income designations (Domina et al., 2018; Fazlul, Koedel, and Parsons, 2021). However, the research literature on the measurement properties of FRM data is thin, and we are not aware of any public policy documents that express

[2] For example, as of 2021, 44 states allocated at least some funds to school districts based on the enrollment of "low-income" students (Source: Education Commission of the States at link (retrieved 06.20.2022): https://reports.ecs.org/comparisons/k-12-and-special-education-funding-2021).

[3] Of the 44 states that allocate funding to school districts based on the enrollment of "low-income" students, 33 use FRM enrollment as at least part of the definition of "low income," and 23 use FRM enrollment exclusively (Source: Education Commission of the States at link (retrieved 06.20.2022): https://reports.ecs.org/comparisons/k-12-and-special-education-funding-2021).

an urgency to understand the implications of using FRM enrollment to inform consequential education policies.[4,5]

In addition to the dearth of research on the measurement properties of existing measures, there has also been little innovation in the field of risk measurement, at least with respect to risk measures for use in education policies. Most policy measures have not changed meaningfully since the 20th century, despite substantial gains in our capacity to collect and analyze data. The long-standing use of a handful of blunt categories, as in current systems, is only preferable to a more holistic, data-driven approach if there is no marginal information to be extracted from the bevy of information about student risk available in state data systems. 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).

There are two possible directions of research motivated by the current state of affairs in risk measurement. One is to expand our understanding of the properties of existing risk measures and the implications of using different measures in different education policies and contexts. The other is to develop new risk measures that address some of the limitations of existing measures in order to give policymakers (and researchers) more options and a broader understanding of what we can measure about student risk using modern data systems and analytic tools. Our contribution is along the lines of the latter.

3. A Framework for Constructing PAP

PAP measures the risk a student faces of poor academic performance. Academic performance can mean many things—achievement on standardized assessments, on-time grade

[4] It wasn't until the introduction of the community eligibility provision (CEP) to the NSLP in 2015 that policymakers began to seek out alternative income indicators, the most popular of which is direct certification (DC) status (Chingos, 2016; Greenberg, 2018). A plausible explanation is that the CEP changed the data in a highly visible and salient way. While there is reason to believe DC status is an improvement over FRM enrollment as a measure of family income (Chingos, 2018; Fazlul, Koedel, and Parsons, 2021; Greenberg, 2018), like with FRM enrollment, there is very little rigorous research on the properties of DC data and the implications of using DC status as a proxy for family income in education policies.

[5] In addition, measurement problems are not unique to income-based risk indicators. Other common risk indicators are also subject to concerns about measurement accuracy but have received less attention in research, in part because it is harder to quantify their limitations. For discussions and evidence on ELL and IEP misclassifications see Abedi (2004, 2008), Sullivan (2011), and Winters, Carpenter II, and Clayton (2017).

progression, school attendance, high-school graduation, college attendance, etc.—and in principle, PAP can be built around any of these concepts. However, for the bulk of our presentation, we anchor PAP to achievement on state assessments. State assessments are the most widespread and differentiated indicators of academic performance available in the education system, and research causally links test scores to consequential later life outcomes such as college attendance and earnings.[6] In the extensions section below, we also consider the potential for using alternative measures of academic performance to anchor PAP.

Moving forward with a test-based anchor in mind, the foundation of our framework is a predictive linear regression of student test scores on student attributes, which can be expressed in broad terms 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 discuss the considerations in selecting the variables for $\mathbf{X_i}$, and how we specify $\mathbf{X_i}$ in our empirical application using the Missouri SLDS, below.

It is also worth noting that our decision to use a simple linear model is with an eye toward feasible policy use. As an academic exercise, it would be useful to extend this basic prediction framework using more complex methods to see if the quality of predictions can be meaningfully improved. However, we view this as a refinement of the framework rather than a core feature. Moreover, our findings suggest that the quality of predictions plateaus quickly conditional on the variables that are generally available in state data systems, which suggests that the likely gains in

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

predictive accuracy from increasing model complexity will be small (in particular, see Table 2 and the corresponding discussion below).

Moving onto model output, 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 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 follows: If the aim is to define risk status based on S_i, why bother estimating $\hat{S}_i$ when S_i observed? There are two reasons, one practical and one conceptual. The practical reason is that S_i will be missing for some students—e.g., untested students when S_i is a test score. However, values of $\hat{S}_i$ can be calculated for all students as long as we observe $\mathbf{X_i}$. The extrapolation to students with missing S_i, such as those who miss tests or are 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. Below we 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.

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*. This is appealing from an incentive-design perspective. To understand why, consider the intended use of the risk measures we aim to develop, which is to inform consequential state funding and accountability policies. Given this intended use, a principle of our framework is that the risk measures 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 performance, S_i. This would perversely incentivize the production of lower test scores (and even if educational actors ignored their perverse incentives, schools and districts with lower test scores would still receive more funding, which would feed into the perception that poor performance is rewarded). However, if risk status is defined by $\hat{S}_i$—that is, by how students are predicted to perform based on their attributes given statewide performance patterns—this undesirable design feature disappears.

A similar logic applies to the selection of the predictive attributes in the **X** vector. These variables should be chosen so that they cannot be manipulated by educational actors to the extent possible. We refer to this as the "non-manipulability" principle of the PAP framework.

4. Empirical Application

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. The foundation of the PAP framework is the cross-sectional regression described in broad terms by equation (1), which we estimate using the 2016-17 (hereafter: 2017) student cohort in Missouri. Before getting into the details of our empirical application, we remind the reader that this is a proof-of-concept exercise. Our goals are to (a) determine the feasibility of constructing PAP as a risk metric and (b) assess its potential value for use in policy. To do this, we must specify the precise variables that we will include in equation (1), how they will be constructed, etc. Although we try to make reasonable choices at each step of implementation, our proof-of-concept exercise is not meant to be prescriptive with regard to exactly how PAP should be constructed. Implementation in other contexts can deviate from our implementation depending on a wide range of factors including data availability, political and legal constraints, policymaker priorities, etc. With this caveat in mind, we proceed with the details of our construction of PAP.

Equation (1) identifies two major components of the prediction framework: outcomes and

predictors. As noted above, for outcomes we use student test scores—specifically, on state assessments in math and ELA in grades 3-8 in Missouri. 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 used 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 apply the prediction model out-of-sample to untested students—i.e., we assume the model would predict their scores accurately if they were tested, on average.[7] Under this assumption, we produce values of $\hat{S}_i$ for all students in Missouri in grades 3-12.[8]

Next, we turn to the predictors of academic performance. We include 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 list of available variables is in Table 1 and includes 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. We construct the panel variables as three-year averages of the individual-student variables taken over the current

[7] We cannot directly test the assumption that the predictions are the same for tested and untested students. However, we can test whether $\hat{S}_i$ is sensitive to using different combinations of tested grades to estimate equation (1). If we obtain substantively different predictions depending on whether we estimate the model on students in grades 3-8, versus grades 3-5, versus grades 6-8, it would suggest the predictions are sensitive to the estimation sample. Such a finding would be inconsistent with the assumption required to use out-of-sample predictions for untested students. Alternatively, if we obtain similar predictions as the test-taking sample changes, it would suggest that the predictions are generally stable. In Appendix Table A1, we show that values of $\hat{S}_i$ from versions of equation (1) estimated using different subsets of tested grades are highly correlated, suggesting it is reasonable to apply the predicted values out-of-sample to untested students.

[8] 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 described below given that students' data histories do not begin until kindergarten—but in principle the predictions can be extended to earlier grades.

and two preceding years.[9,10] These variables capture the persistence of students' circumstances and are motivated by prior work on the predictive validity over academic outcomes of persistent poverty (Michelmore and Dynarski, 2017) and mobility (Goldhaber et al., 2022). The third 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.

Again, we do not wish to prescribe the precise set of variables that should be used in equation (1) or how the variables should be constructed. There are legal, political, and other considerations that will guide these decisions in different contexts (we elaborate on one example—the exclusion of racial/ethnic categories—in more detail in section 6). However, in order to proceed with our empirical application, we must implement decision rules that allow us to construct PAP.

4.2 Practical Issues

4.2.1 Excluded Variables and Adherence to the Non-Manipulability Principle

In the ideal implementation of our framework, the variables in the **X**-vector would be non-manipulable. The non-manipulability principle leads us to exclude some types of information 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; Therriault et al., 2017).

---

[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] Our use of three-year averages (as opposed to, say, count variables) allows us to use three years of data for students 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-enrolled in both years is coded as "100 percent" FRM-enrolled, 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 would be treated as having incomplete **X** vectors). However, this is an inferior option from a policy perspective because we can estimate $\hat{S}_i$ for fewer students.

However, although these variables are useful in diagnostic applications, they are a poor fit for PAP because they can be affected (potentially a great deal) by district and school behavior. 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 (Bass, 2010). Schools and districts can also potentially manipulate other student categories including ELL and IEP status. 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. 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 prediction variables we use. The tradeoff between non-manipulability and predictive accuracy merits consideration in any policy application of our framework (or any other risk measurement framework).[12]

[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 for future research.

[12] Another variable-selection issue is whether to use information external to the education system in the prediction model. All else equal, using internal data is desirable because external data can be altered in ways the education system cannot control, such as occurred when the NSLP implemented community eligibility. However, an application of our framework that is entirely internal to the education system would not include any family-income indicators (both FRM and DC data are from external programs). As we show below, this has significant adverse consequences for the predictive accuracy of equation (1). Ultimately, we decided to include DC status as an external indicator of poverty in our implementation of PAP. This is again in the spirit of striking a balance between competing priorities. This choice is made easier by evidence we show below that student risk designations based on our framework are far less sensitive to data disruptions (e.g., the Community Eligibility Provision) than risk designations in common categorical systems.

4.2.2 Variable Weights

Even if the elements of **X** are selected to be non-manipulable, the weights—contained by the vector $\boldsymbol{\beta}_1$ in equation (1)—can still be influenced by school demographics and policies. The weights could be problematic if a particular district enrolls a disproportionate share of students with one or more of the predictive attributes. In that case, the district's own performance could meaningfully influence the weights on those 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.

Fortunately, this concern can be circumvented by jackknifing, which is an estimation procedure that prevents individual schools and districts from influencing their own weights. 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.[13] 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*, **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 faced by students in district *j* based on their attributes, as predicted by a statewide model outside of 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 in academic research, but at least in its purest form, it can be computationally intensive and may be unnecessarily complex for policy applications. Therefore, we explore the use of simpler variants of the jackknifing procedure. Our preferred

[13] 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).

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 Appendix A, 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.

4.2.3 Risk Status Indicators

We use the values of $\hat{S}_i$ to divide students into high-risk and low-risk categories, mirroring the categorical structure of existing risk measurement systems in education. This facilitates apples-to-apples comparisons of PAP to other competing measures of risk. Of course, an advantage of PAP over other risk metrics is that the underlying predicted values, $\hat{S}_i$, contain more differentiated information about risk than is reflected by the binary categories—we return to this point in the extension section below. For now, we divide students into risk categories by specifying a threshold test value, $\tilde{S}$, that separates low-risk and high-risk students. We choose $\tilde{S}$ based on predicted test proficiency to align the categories with test policies, which are often proficiency-based, although this is not a prescriptive aspect of PAP.

A full-scale policy implementation would likely match $\tilde{S}$ to proficiency targets on state tests, which are grade and subject specific. Mirroring this approach, but simplifying it and creating a degree of separation from the specific policy context in Missouri, we set a single threshold value of $\tilde{S}$ for all grades and subjects based on 2017 NAEP performance. Specifically, 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, and we use this percentile threshold—the 26.25th percentile—as $\tilde{S}$. We then apply this threshold to 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 on the MAP as high-risk students, and students above the 26.25th percentile as low-risk students.

Our simplified approach to setting $\tilde{S}$ based on NAEP data does not have any substantive bearing on how our framework operates. 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.[14]

4.3 Statistical Summary of PAP

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 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 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 variable interactions. The notes to Table 2 give precise details about each specification.

The columns provide statistical information about the models. In column (1), the R-squared values range from 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.

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

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 but are not accounted for in the model. 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.[15] 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 about 0.39.[16] 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 about the efficacy of our predictions from this number. 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 value of PAP based on the predictive power of the model.

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 score below $\tilde{S}$. 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 assignments based on their actual test scores. A false-positive is a

[15] 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.

[16] The effective R-squared can be obtained by dividing the estimated R-squared by the maximum value (Aaronson, Barrow, and Sander, 2007).

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, the table does provide several useful insights. One is that poverty status variables—whether FRM or DC—add substantial predictive value to the model. Relative to our specifications in rows (*a*), (*e*), and (*i*) that omit this information, the R-squared values increase by 3-5 percentage points in rows where we include it in various forms.[17]

Another insight 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 or additional interactions, although as noted above, future work could examine the potential to improve predictive accuracy via modeling and estimation adjustments more formally.

Table 3 summarizes our risk measures overall, and within traditional categories of

[17] Between the two, FRM data are more predictive of test scores than DC data (this can be seen by comparing 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 explanatory power to the model because there is more variance in FRM data (i.e., the FRM-enrolled student share is closer to 0.50). 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.

disadvantage, by reporting means and standard deviations of $\hat{S}_i$ from our preferred specification in row (*l*) of Table 2.[18] 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 help to motivate the policy use of traditional categories of disadvantage. The gaps in average predicted achievement by DC status, FRM status, ELL status, 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.[19]

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 information is in the second row of the table, which reveals substantial 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.[20] 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-enrolled 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 useful descriptive information, they are not directly informative about the utility of PAP. This is because they do not address the policy-relevant question of whether PAP is "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

[18] None of the substantive findings in Table 3 are unique to using our preferred specification.
[19] We do not report values for the many coefficients from our prediction models because the multivariate regression framework makes their interpretation 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.
[20] 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.

these dimensions into our empirical application by using different definitions of student risk to implement simulated funding policies.

4.4 Policy Simulations

We simulate a weighted student funding formula. In the simulations, every student receives a foundational per-pupil amount, and students designated as "high risk" are allocated additional funding on top of the foundational amount. We allocate a fixed education budget according to the formula and compare the resulting allocations when we use different metrics to identify high-risk students. We briefly summarize the simulation design and results in this section and refer interested readers to Appendix C for additional details and full results.

We consider systems that identify high-risk students using (1) DC status, (2) FRM status, and (3) PAP status. We also simulate California's LCFF as a real-world policy comparison using the Missouri data—we refer to the Missouri-based implementation of the LCFF as the "pseudo-LCFF." The pseudo-LCFF incorporates both FRM and ELL as risk indicators, and also includes a "concentration" component that directs additional funding to districts with concentrated risk as measured by the unduplicated share of ELL and FRM students (this is based on the true policy implementation in California—see Appendix C for details).

The simulations show that when we designate students as high-risk based on DC or FRM status, the funding formula allocates more resources to DC or FRM students, respectively, than when we use PAP to designate risk status. Similarly, the PAP-based system allocates more resources to students with lower predicted test scores. This is all by design. In addition, we show that a system based on PAP allocates more funding to students with lower *actual* test scores. This may seem trivial—and in some sense it is—but it is notable that PAP allows us to compare allocations between students with low predicted and actual scores. An analogous calculation in the poverty-based systems is not possible because state education agencies do not have data on actual incomes. Finally, the PAP-based system allocates more funding than analogous DC- and FRM-based systems to ELL, IEP, and URM students, and often by a sizeable margin. This is because PAP takes explicit account of these risk categories (through their associations with

student test scores), whereas the DC- and FRM-based systems do not.

Next, we turn to our comparisons to the pseudo-LCFF. We find that along every dimension of risk that we measure save one, the PAP-based system allocates more funding to high-risk students. The exception is FRM students, who are allocated more funding by the pseudo-LCFF. This is not surprising because the pseudo-LCFF directly targets FRM students.[21] However, it is notable that among the risk categories better targeted by the PAP-based system is DC status, which like FRM status also measures student poverty, and does so more accurately (Fazlul, Koedel, and Parsons, 2021).

The PAP-based system is better at targeting resources to high-risk students along most dimensions than the pseudo-LCFF. There are two reasons for this. First, it explicitly accounts for more dimensions of risk via the test prediction model. Second, the funds targeted for high-risk students are less diluted in the PAP-based system, accruing to the bottom 26.25 percent of students. In contrast, under the pseudo-LCFF, 51.1 percent of Missouri students are identified as high risk (the unduplicated sum of FRM and ELL students), spreading the supplemental funds available to support them more thinly under the fixed budget.

In summary, our policy simulations incorporate realistic counterfactuals and show that using PAP to identify high-risk students in a funding policy would meaningfully change the allocation of resources. In terms of targeting students with low predicted or actual academic performance, a PAP-based system is clearly preferred. Along other dimensions of measured risk, there are tradeoffs that depend on policy objectives. We conclude by noting our simulations only consider systems that fully substitute between risk metrics. PAP could also be used to augment existing systems as an additional dimension of funding consideration.

[21] The pseudo-LCFF also directly targets ELL students, but it does not allocate more funding to ELL students compared to the PAP-based system in Missouri. However, we do not emphasize this finding in Missouri because the ELL share is small and it may be idiosyncratic. See Appendix C for details.

5. PAP Flexibility in Response to Changes in the Underlying Data

An advantage of PAP 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 the meaning of this variable were to change in a way that makes it less informative, perhaps due to a change to the state's Broad Based Categorical Eligibility policy, it could greatly influence measured risk.[22] In contrast, the effect of the same data change on PAP will be dulled because of the remaining predictors in the **X** vector. That is, to the extent the other predictors 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 4 we report correlations of $\hat{S}_i$ as estimated by all of the specifications shown in Table 2. The correlations are reported in reverse order for the specifications in rows (*l*) to (*a*) of Table 2 in order to emphasize differences among our preferred models. Column 1 shows that compared to our primary specification in row (*l*), most alternative specifications yield similar risk values for students. For example, only 3 of 11 correlations in the first column are below 0.89, and the minimum value is 0.84 from row (*a*), which is the sparsest specification. In fact, the correlation reported between models (*a*) and (*l*) of 0.84 is the minimum value in the entire correlation matrix. Broadly speaking, Table 4 confirms that the risk metrics for individual students are fairly stable as we change the attributes included in **X**.

Next, in Appendix D we assess the implications of a hypothetical switch from using *free meal* (FM) status to using DC status to identify students from low-income families. A similar data substitution—i.e., from using FRM status to using DC status—is occurring or under consideration in many states today due to concerns about the accuracy of FRM data with the

[22] Broad-based categorical eligibility policies allow families with higher incomes to qualify for SNAP, the primary program that is used for direct certification.

introduction of community eligibility for free meals in the NSLP (Chingos, 2018; Greenberg, 2018). The data switch we consider is conceptually more appealing because FM enrollment and DC status in Missouri are purported to identify students from families at the same income threshold—130 percent of the poverty line or below.[23]

The results in Appendix D make clear that the flexibility of our approach is a significant practical benefit in the event of this data substitution, as PAP is much less volatile than the categorical alternative. Specifically, using PAP, 4.4 percent of students experience a change in their risk status due to the change from using FM to DC data to identify students from low-income families, compared to 17.8 percent of students using the categorical system. From a statistical standpoint, the model-based approach that undergirds PAP must perform better than the categorical approach, and in this sense the comparison we report in Appendix D is somewhat of a straw man. However, it is important to recognize the categorical comparison is a fundamentally accurate characterization of current systems, which helps to explain the policy consternation as states consider new definitions of low-income status in reaction to the introduction of community eligibility in the NSLP.

6. Extensions & Policy Considerations

6.1 PAP Without Taking Explicit Account of Race-Ethnicity

The goal of the models summarized by Table 2 is to predict academic performance, and race-ethnicity is a consistently strong predictor. Statistically, whether it is desirable to use information on race-ethnicity to improve the predictions is unambiguous: these data should be used. However, some have argued that race-ethnicity data should be omitted from prediction models such as ours based on the concern that it sets different expectations for academic performance across racial-ethnic groups.[24] Moreover, depending on the intended use of PAP, it may not be legal to allow for a direct predictive impact of race-ethnicity. As such, we briefly

[23] Although in practice FM enrollment is oversubscribed, partly due to community eligibility and partly for other reasons, as shown by Domina et al. (2018) and Fazlul, Koedel, and Parsons (2021).
[24] Ehlert et al. (2016) provide a deeper discussion on this issue in the context of school accountability systems.

consider how the omission of racial-ethnic information from the prediction model impacts PAP.

First, 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 race-ethnicity (though, of course, $\hat{S}_i$ will still be correlated with race-ethnicity given the inclusion of other predictors that are correlated with both race-ethnicity and student outcomes). Appendix Table E1 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 E2 we use our policy simulation to show how 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 Appendix Table C1. Most of the findings are similar regardless of whether we use the full or restricted versions of model (*l*), which is as expected given the general robustness of the prediction framework shown in Table 4. 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 (still, the amount allocated to URM students is larger than in funding formulas that rely on FRM or DC data, or the pseudo-LCFF).

6.2 Augmentation for IEP Students

It is also worth noting that there are some aspects of current systems that PAP does not meaningfully improve upon. The most obvious example is students with severe IEPs, for whom both broad categorical designations and PAP are insufficient to capture the extent of their needs. As a result, funding add-ons will be needed for IEP students to augment any general framework. For accountability policies, which we discuss briefly below, it may also be desirable to exclude

these students, or at least those whose disabilities are deemed sufficiently severe, as is already common practice in many states.

6.3 Using Other Academic Outcomes to Anchor PAP

There is nothing that requires PAP to be anchored to student test scores. In this section we consider two alternative academic performance indicators—student attendance and high school graduation. Each of these possibilities has strengths and weaknesses compared to test scores. Our view is that test scores offer the best combination of properties, but again, the use of test scores as the anchor for PAP is not a prescriptive feature.

First, student attendance is linked to a broad swath of indicators of disadvantage (Ready, 2010). Of the available alternatives, it is also the one with the greatest data coverage—virtually every student should have an attendance record in every year. However, there are two main limitations of using attendance to anchor PAP. First, conceptually, attendance is not a pure measure of academic performance—it is arguably more of an input than outcome—and may be undesirable for this reason. Second, student attendance is a less-differentiated measure than student achievement on state tests, and relatedly, the data elements available in state data systems do not predict variation in attendance as well as they predict variation in test scores. For example, we estimated a version of model (l) from Table 2 where the attendance rate was the dependent variable, and the R-squared was just 0.067, compared to 0.29 for test scores.[25] These technical limitations will lead to less differentiation in PAP anchored to attendance.

High school graduation is another interesting alternative. It is appealing conceptually because graduation is an important goal of the education system. However, as an anchor for PAP, it is problematic for several reasons. First, graduation rates can be improved by increasing learning and improving student supports *or* by reducing standards, and it is difficult to disentangle these mechanisms. If different standards are applied to students who differ by their

[25] One way to quickly convey the (relatively) limited variation in attendance is to note that a large fraction of students have very high attendance—e.g., in Missouri in 2017, 87 percent of students attended 90 percent of days or more).

X-vector attributes, PAP would produce misleading risk gaps. Test scores do not have this limitation because all students take the same test. Another challenge of using high school graduation is that only 12th grade students graduate each year (with few exceptions). This means that PAP will always be backward looking for most students—i.e., for each cohort below grade-12, PAP will assign risk values based on how their attributes predicted graduation for an older cohort, and potentially a much older cohort (for early-grade students). Finally, a third limitation of using high school graduation is that it is a relatively undifferentiated measure of academic performance. It splits students into just two blunt categories (graduated or not), and most students graduate (the national graduation rate is well above 80 percent).

In the context of these alternatives, test scores are appealing given their combination of (a) data coverage (a large fraction of the student population is tested), (b) both differentiation and predictable differentiation, and (c) conceptual alignment with one key purpose of schooling, which is to promote student learning. That said, future research could consider combining PAP anchored to test scores and other student outcomes, including but not limited to attendance and high school graduation, in order to provide a more holistic indication of risk along multiple dimensions of academic performance.

6.4 Monitoring Achievement Gaps

The policy simulations discussed above and presented in Appendix C focus on school funding. However, PAP also has features that make it appealing for use in other policies, such as for monitoring achievement gaps within schools.[26] PAP can consolidate accountability information across the multiple categories of risk tracked by many states (e.g., FRM, ELL, IEP, URM) into a singular indicator, which in turn can reduce information overload and mitigate type-I errors resulting from the multi-category approach (Davidson et al., 2015). Some states do something similar now with "super subgroups" that combine students from multiple risk categories, but current practice does not account for compositional differences in the super

[26] Many states informally monitor within-school achievement gaps, and these gaps are incorporated into some states' formal accountability policies (Martin, Sargrad, and Batel, 2016).

subgroup across schools, clouding inference. Using PAP for accountability offers the simplifying benefit of the super-subgroup approach, while at the same time minimizing the potential for misleading comparisons due to differences in the composition of super subgroup across schools. We elaborate on the use of PAP in accountability policies in Appendix F.

6.5 Uncoarsened PAP

In our policy simulations, we use binary risk categories to group students based on their underlying PAP values. This facilitates a straightforward comparison to *status quo* systems and lends policy relevance to our work given the strong cultural norm in education of grouping students categorically (and often in a binary fashion). However, by coarsening $\hat{S}_i$, we strip away much of the differentiating information about student risk it contains.

In the interest of brevity, we do not examine the potential for using 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 three-, four-, or 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.[27]

7. Conclusion

We develop and test a new measure of student risk, which we call Predicted Academic Performance, or PAP. PAP measures student risk in precise terms—it captures a student's risk of poor academic performance. It also modernizes the approach to risk measurement compared to available alternatives. PAP is a flexible measure 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, PAP is designed for use in

[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.

consequential education policies. 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.

Our primary contribution is to put 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 risk measurement practices in education policy. This is despite the fact that 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 the policy value of PAP using the Missouri SLDS in a proof-of-concept exercise. 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 these precise details of constructing PAP. In fact, some of our findings refute the notion that there is a clear "right way" to estimate PAP that would merit a prescriptive recommendation, which we view as a desirable feature of the framework.

Once implemented, PAP 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 Missouri data suggest that PAP will not change dramatically in response to most data changes. Moreover, some "drift" in PAP from year-to-year as new measures become available, and/or relationships between students' underlying risk indicators and academic performance change, may be desirable as it will allow PAP to adjust over time to evolving education circumstances. The alternative is a stagnant measurement system that changes less often but more disruptively.

Variants of PAP could also measure risk along other dimensions of academic performance—e.g., in terms of attendance, graduation, and college matriculation. These could replace test scores in the framework or, more likely, augment them. For instance, each student's total risk score could be a weighted average of risk assessed by different indicators of academic performance. PAP could also be applied to 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.

The impetus for the development of our framework is the inadequacy of the measures of student risk currently used in consequential educational policies. No new approach, including ours, will perfectly measure risk due to the inherent difficulty of the task. But despite their limitations, new systems can complement and 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.

Abedi, J. (2008). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice* 27(3).

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher* 33(1), 4-14.

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.

Chingos, M.M. (2016). No more free lunch for education policymakers and researchers. *Evidence Speaks Reports* 1(20), 1-4. 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 enrollment 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.

Harwell, M., & LeBeau, B. (2010). Student eligibility for a free lunch as an SES measure in education research. Educational researcher, 39(2), 120-131.

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.

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

Ready, D.D. (2010). Socioeconomic disadvantage, school attendance, and early cognitive development: The differential effects of school exposure. *Sociology of Education* 83(4), 271-286.

Sullivan, A.L. (2011). Disproportionality in special education identification and placement of English language learners. *Exceptional Children* 77(3), 317-334.

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.

Therriault, S.B., O'Cummings, M., Heppen, J., Yerhot, L. and Scala, J. (2017). Early warning intervention and monitoring system implementation guide. Lansing, MI: Michigan Department of Education.

Winters, M.A., Carpenter II, D.M., and Clayton, G. (2017). Does attending a charter school reduce the likelihood of being placed into special education? Evidence from Denver, Colorado. *Educational Evaluation and Policy Analysis* 39(3), 448-463.

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 Enrolled	0.50	0.50
Free-Lunch Enrolled	*0.44*	*0.50*
Reduced-Price Lunch Enrolled	*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. 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.

Appendices
(Online Only)

Appendix A: Supplementary Tables

Appendix Table A1. 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] Instead, in Figure B1 we show the distributions of predicted scores, $\hat{S}_i$, for the different specifications shown in Table 2 in the main text.

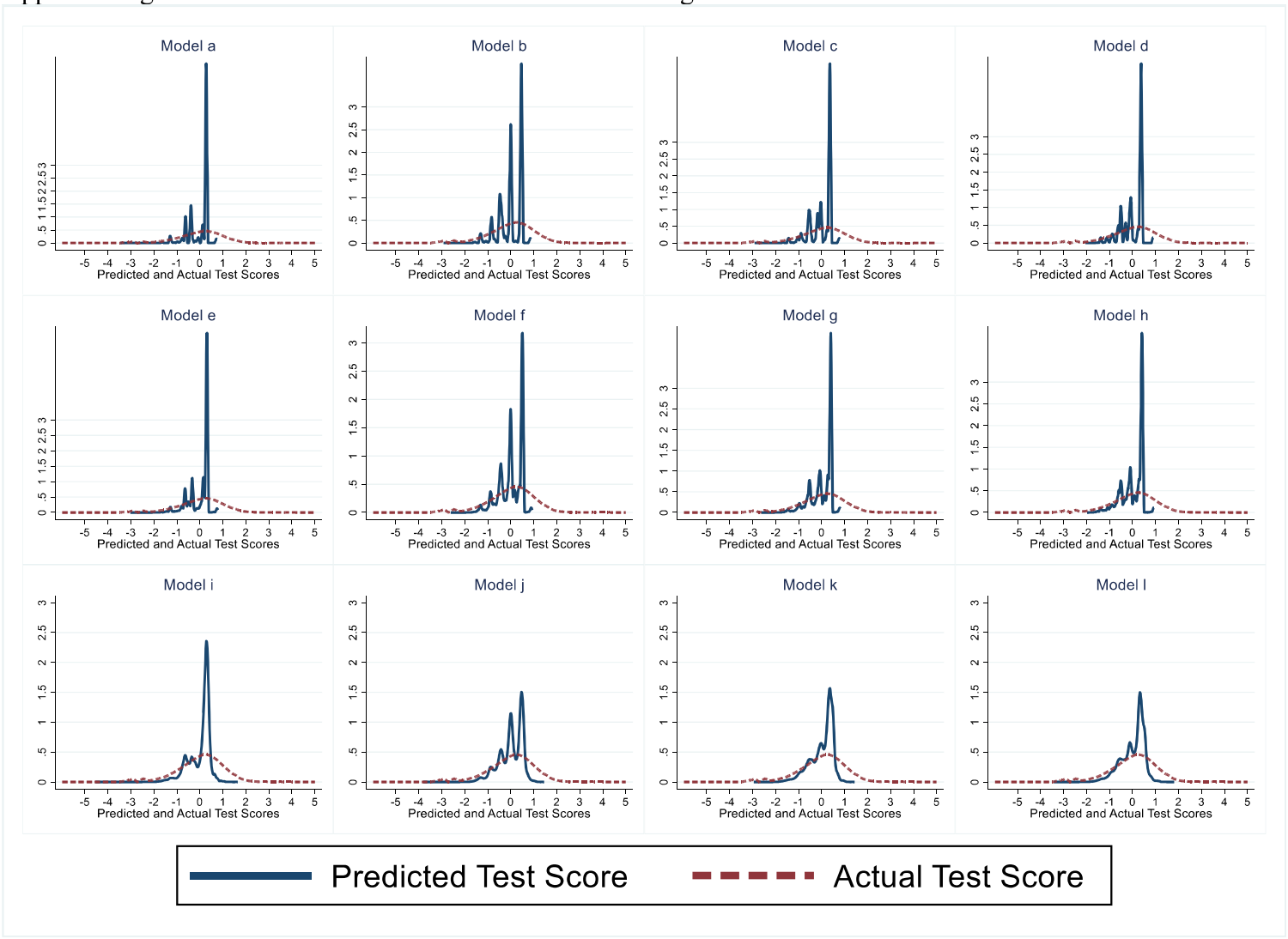
There are two reasons we show the distributions. The first is to highlight their "lumpiness," especially for the sparser versions of the prediction model. The lumpiness is not surprising because all of the student-level control variables in the models are binary or categorical indicators. When we add 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 columns 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.

The lumpiness 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 permit 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: Funding Policy Simulations

To truly understand the potential value of PAP, it must be examined within a framework that accounts for the next-best alternatives. In this appendix, we report on our policy simulations that illustrate how using PAP in a basic student-weighted funding formula would affect the allocation of resources to different types of students and schools compared to common alternatives. The funding formula we use to allocate resources to "high risk" and "low risk" students in the simulations is as follows:

$$N_L + (1+Z)N_H = B \qquad \text{(C1)}$$

In equation (C1), N_L is the number of low-risk students, N_H is the number of high-risk students, and *B* is the total budget. The amount allocated to each low-risk student is normalized to 1.0, and *Z* is a positive multiplier that captures the additional per-pupil resources 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.

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 (C1):

$$Z = \frac{B - N_L}{N_H} - 1 \qquad \text{(C2)}$$

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.

Appendix Table C1 shows results from our first set of policy simulations. We use the risk measures from our framework to allocate resources to students via equation (C1) and compare

the allocations to alternative allocations based on 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- and low-risk students.

Each column of Table C1 shows results from a different policy parameterization, defined by the first four rows. The subsequent rows 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 0.263 in the table). From equation (C2), with *B = 1.25N*, the third row of the table shows that *Z* is *0.952*. Thus, the policy allocates 1.952 resource units to each high-risk student and 1.0 resource units to each low-risk student.

The bottom panel of the table shows 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 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 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 performance (i.e., $\hat{S}_i$).

The normalization of resource units facilitates straightforward comparisons within and across columns in the table. The easiest way to compare the allocations is in percentage units

relative to the normalized baseline allocation of 1.0, which in a funding system would correspond to a foundational per-pupil dollar value. 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 the first columns for each of 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 or enrolled in the NSLP, respectively, so 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. The second columns for these scenarios define DC and FRM students as at-risk students. This allows for comparisons of PAP to the alternatives 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. The finding for FRM students follows from the strong overlap between FRM enrollment and DC status. However, the targeting of 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, the PAP-based system 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 PAP is more effective at identifying students at risk of poor academic performance than DC status. The PAP-based system 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 allocations are smaller compared to Scenario 2 because *Z* is

smaller (which, in turn, is because there are so many FRM-enrolled students), which suppresses the per-student allocations under the fixed budget. Still, the general pattern of findings holds. 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 PAP against the FRM-based alternative inclusive of the difference in the value of *Z*. The comparison shows PAP allocates more resources per student along every measured dimension of risk except FRM status (including DC status, albeit marginally). For most non-FRM characteristics, PAP 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 (*Z*). A summary of this comparison is as follows: the PAP-based policy is more effective at targeting resources toward high-risk students both (a) as identified in terms of academic performance and (b) using most other common categorical definitions.

The comparisons in Table C1 are informative but generic. In Appendix Table C2, we make 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 allocations based on PAP 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.[29] We implement a modified version of LCFF that ignores foster youth because we do not have access to data for this designation in Missouri.[30]

[29] 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).

[30] 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-enrolled.

The LCFF also provides additional per-pupil funding to districts with concentrated need (Johnson and Tanner, 2018). Based on 2021 LCFF funding rules, we convert the LCFF formula to a student-level allocation model to fit within our analytic framework. The student-level version of the LCFF funding formula is as follows:

$$F_i = F_0 + (0.2 * F_0) * D_i + (0.65 * F_0) * \max[D_d - 0.55, 0] \quad \text{(C3)}$$

In equation (C3), 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, 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. Following our analytic structure from above, we normalize F_0 to 1.0.[31]

We apply the "pseudo-LCFF" in Missouri and assign each student a value for F_i. To compare the subsequent student allocations to PAP-based allocations, 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 (C1). For notational convenience, we write the total budget in units of *N* as above (applying the LCFF rules in Missouri, B=1.152N). Then, using this budget, we allocate resources following equation (C1) and set $\tilde{S}$ at the 26.25th percentile of test scores. This facilitates a fixed-budget comparison between the pseudo-LCFF and the PAP-based alternative.

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" component of the formula allocates more resources to districts with concentrated need. Our PAP-based policy structure does not include a directly-analogous

[31] The way the student-level formula is written in equation (C3), each student in a district with $D_d > 0.55$ receives a small positive increment, which is mathematically equivalent to identifying the fraction of students above 55 percent and providing the district with the full increment for each of these students.

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 lower values of $\hat{S}_i$ to students in high-concentration schools. (We could also modify our policy structure to copy the LCFF by allocating more resources to schools and districts with high proportions of low-$\hat{S}_i$ students, although we do not pursue this extension here.)

Table C2 shows the results comparing resource allocations based on PAP to the pseudo-LCFF in Missouri. Along most dimensions, including the key metrics of test performance and predicted test performance, PAP 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 modestly higher resources. However, notably, this result does not translate to DC status under LCFF, which is a more accurate measure of poverty (Fazlul, Koedel, and Parsons, 2021).[32]

PAP yields higher per-student allocations along most dimensions because it explicitly accounts for them in the test prediction model. Moreover, the funds targeted for high-risk students are less diluted under PAP, accruing to 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 C1 and C2 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 resources will impact school-level allocations, whether in our framework or 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 because each school's student body

[32] 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).

would be of the same proportions (subject to sampling variance). In the real world, however, residential sorting implies that changes to student-level allocations will translate at least partly to changes in school-level allocations.

Appendix Tables C3 and C4 provide information complementary to Tables C1 and C2, but at the school level. This allows us to assess the extent to which the resource differences across students shown in Tables C1 and C2 translate to cross-school differences, in acknowledgment of the difficulty of targeting resources to individual students differentially within a school.

We start with Table C3, which replicates the scenarios from the initial policy simulations shown in Table C1. Reflecting this, the first four rows of Table C1 and Table C3 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 Table C3 are in the expected direction following on Table C1, although the magnitudes of the correlations vary depending on how students are distributed across schools.

We highlight two key takeaways from Table C3. First, the tautological aspects of the allocations from Table C1 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 C1, 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).

Table C4 reports the same school-level correlations under the pseudo-LCFF. Again, the general insights from the student-level resource allocations in Table C2 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.

Appendix Table C1. 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	0.263	0.273	0.273	0.503	0.503
N(L) Share	0.737	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.

Appendix Table C2. 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	0.263	0.511
N(L) Share	0.737	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).

Appendix Table C3. 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	0.263	0.273	0.273	0.503	0.503
N(L) Share	0.737	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 C4. 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	0.263	0.511
N(L) Share	0.737	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 D: Implications of a Concrete Data Change for Student At-Risk Designations

In Appendix Table D1, we assess the implications of a hypothetical switch from using *free meal* (FM) status to using DC 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 enrollment is oversubscribed—see Domina et al., 2018; Fazlul, Koedel, and Parsons, 2021). In the categorical system, we recode students as high risk based on DC status instead of FM status. For PAP, we make the same data switch in the prediction model. For the DC-data scenario we estimate the model 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 high-risk students in our framework based on predicted achievement—i.e., a high-risk student has $\hat{S}_i < \tilde{S}$, where $\tilde{S}$ is set at the 26.25th percentile.

The results in Appendix Table D1 make clear that the flexibility of our approach is a significant practical benefit. PAP is much less volatile than the categorical alternative in response to the hypothetical data switch. Specifically, the data switch results in just 4.4 percent of students switching at-risk status as measured by PAP, compared to 17.8 percent of students under the categorical alternative.

The reason for the difference is that the weighting parameters in the prediction model adjust to reflect the informational content of the new variable, holding the share of at-risk students fixed. We largely identify the same group of at-risk students regardless of whether we use FM or DC data. The change in the categorical designations is much more disruptive because of the large difference in the size of the FM and DC categories and the inherent inflexibility of the categorical approach.

Appendix Table D1. Changes in Student Categorizations as At Risk in the Hypothetical Condition where DC Data are Used in Place of FM data to Identify Students from Low-Income Families.

	PAP-Based Framework: $\hat{S}_i$ is predicted with DC data using the model shown in row (*l*) of Table 2, and again 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 FM status, then by DC status
Share of high risk students using FM	0.263	0.441
Share of high risk students using DC	0.263	0.273
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 DC scenario within our framework corresponds to row (*l*) of Table 2; the FM scenario is identical except we replace any DC-based information with FM-based information in the prediction model.

Appendix E: Omitting Information about Race-Ethnicity from the Prediction Model

In this appendix, we briefly report on our findings if we omit information about student race-ethnicity entirely from the prediction model. There is not a strong statistical rationale for omitting race-ethnicity information from the model. Nonetheless, we provide this analysis for completeness and in recognition of the fact that there may be political, legal, or other reasons for its exclusion.

The results from our analysis omitting all racial-ethnic information from the prediction model are provided in Appendix Tables E1 and E2. Table E1 corresponds to Table 2 in the main text, and Table E2 corresponds to Appendix Table C1. In very brief summary, Table E1 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 E2 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. This result follows from Table 4, which shows that using different predictors, and combinations of predictors, generally does not have large effects on 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 E2 compared to Appendix Table C1. 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 high-risk.

Appendix Table E1. 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 E2. Comparison of Primary Policy-Simulation Findings from Appendix Table C1 Using Test Prediction Models With and Without Race-Ethnicity Information.

	Baseline Scenario: $\tilde{S}$ set at basic/below basic achievement percentile (repeated from Table C1)	$\tilde{S}$ 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	0.263	0.263
N(L) Share	0.737	0.737
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 F: Using PAP for Accountability (Within School Achievement Gaps)

In this appendix, we discuss some of the drawbacks of how data are used to track achievement gaps across student groups within schools and describe how PAP can be used to improve upon current practice.

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.

Other problems with the multi-category approach include that it can (a) cause information overload (Sutcliffe and Weick, 2009) and (b) 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 increases statistically (Davidson et al., 2015). 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 belong to the super subgroup. However, its limitation is that there will be compositional differences in the super subgroup across schools, which exacerbates the problem raised in the preceding paragraph of group heterogeneity in expected student performance.

PAP facilitates the single-comparison simplicity of the super-subgroup approach while minimizing the potential for misleading comparisons due to differences in the composition of the 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.

To illustrate, we begin by identifying 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:

$$\frac{1}{N_{L,k}}\sum_{i=1}^{N_{L,k}} S_{ik} - \frac{1}{N_{H,k}}\sum_{i=1}^{N_{H,k}} S_{ik} \qquad \text{(F1)}$$

In equation (F1), 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, $\hat{S}_{ik}$:

$$\frac{1}{N_{L,k}}\sum_{i=1}^{N_{L,k}} \hat{S}_{ik} - \frac{1}{N_{H,k}}\sum_{i=1}^{N_{H,k}} \hat{S}_{ik} \qquad \text{(F2)}$$

The observed and predicted achievement gaps in equations (F1) and (F2) 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 (F1) and (F2):

$$\{\frac{1}{N_{L,k}}\sum_{i=1}^{N_{L,k}} S_{ik} - \frac{1}{N_{H,k}}\sum_{i=1}^{N_{H,k}} S_{ik}\} - \{\frac{1}{N_{L,k}}\sum_{i=1}^{N_{L,k}} \hat{S}_{ik} - \frac{1}{N_{H,k}}\sum_{i=1}^{N_{H,k}} \hat{S}_{ik}\} \qquad \text{(F3)}$$

Momentarily suppressing discussion of one technical caveat, equation (F3) 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 (F3) is negative. Said another way, schools with negative values of equation (F3) have achievement gaps that are smaller than what would be expected based on their compositions of low-risk and high-risk students.

Equation (F3) 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 (F3) 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 is that the fitted values in equation (F2)—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. Due to the shrinkage, the average gap between the test score predictions in equation (F2) will be attenuated relative to the gap in observed scores in equation (F1), resulting in disproportionately positive values for the difference in equation (F3).

Fortunately, as in the allocation-policy context, there are straightforward solutions to address the shrinkage problem. One solution, following from our preceding analysis, is to calculate the values from equation (F3) using percentiles rather than actual and predicted scores. The interpretation of equation (F3) 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 (F3) is calculated 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 (F3) 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.[33] This inflation should

[33] 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 (F3).

occur after the predictions are made using equation (1) in the main text, but before constructing the average predicted values in equation (F2). Using the variance-inflated $\hat{S}_i$ values, equation (F3) can be interpreted in test-based units, and the same inference can be drawn for positive and negative values as described above.