



Do Test Scores Misrepresent Test Results? An Item-by-Item Analysis

Jesse Bruhn

Brown University
and NBER

Michael Gilraine

Simon Fraser University
and NBER

Jens Ludwig

University of Chicago
and NBER

Sendhil Mullainathan

MIT, and NBER

Much of the data collected in education is effectively thrown away. Students answer individual test questions, but administrators and researchers only see aggregate performance. All the item-level data are lost. Ex ante it is not clear this destroys much useful information, since the aggregate might be a sufficient statistic. Using data from Texas for 5 million students and 1.31 billion student-item responses, we show that in fact aggregation does destroy a great deal of valuable information in education: (1) Even conditional on a summary test measure, there is additional information in the item-level data; (2) This additional information is relevant for the student outcomes that education decisions seek to optimize; and (3) This information can be made practically useful for schools. Given how inexpensive storing, transmitting and analyzing such data would be, large gains could be had in education by simply using all the data we currently collect.

VERSION: November 2025

Suggested citation: Bruhn, Jesse, Michael Gilraine, Jens Ludwig, and Sendhil Mullainathan. (2025). Do Test Scores Misrepresent Test Results? An Item-by-Item Analysis. (EdWorkingPaper: 25-1343). Retrieved from Annenberg Institute at Brown University: <https://doi.org/10.26300/npjb-0861>

Do Test Scores Misrepresent Test Results?

An Item-by-Item Analysis*

Jesse Bruhn, Michael Gilraine,
Jens Ludwig, and Sendhil Mullainathan

November 2025

Abstract

Much of the data collected in education is effectively thrown away. Students answer individual test questions, but administrators and researchers only see aggregate performance. All the item-level data are lost. *Ex ante* it is not clear this destroys much useful information, since the aggregate might be a sufficient statistic. Using data from Texas for 5 million students and 1.31 billion student-item responses, we show that in fact aggregation does destroy a great deal of valuable information in education: (1) Even conditional on a summary test measure, there is additional information in the item-level data; (2) This additional information is relevant for the student outcomes that education decisions seek to optimize; and (3) This information can be made practically useful for schools. Given how inexpensive storing, transmitting and analyzing such data would be, large gains could be had in education by simply using all the data we currently collect.

*The conclusions of this research do not necessarily reflect the opinions or official position of the Texas Education Research Center, the Texas Education Agency, the Texas Higher Education Coordinating Board, the Texas Workforce Commission, and the State of Texas. We thank Jiguang Li, Luke Frymire, and Ester Muzychuk for exceptional research assistance, and James Ross for critical research management. For helpful feedback we thank seminar participants at Harvard, the University of Melbourne, the Econometrics Interactions Workshop, the University of Arizona, the Arizona State University Applied Micro Conference, Tufts University, the Federal Reserve Bank of New York, and the NBER summer institute meetings. For financial support, we thank the Student Upward Mobility Initiative, Manny Roman and the Center for Applied AI at the University of Chicago Booth School of Business. All opinions and any errors are ours alone. Jesse Bruhn: Department of Economics, Brown University and NBER. Email: jesse_bruhn@brown.edu. Michael Gilraine: Department of Economics, Simon Fraser University and NBER. Email: gilraine@sfu.ca. Jens Ludwig: Harris School of Public Policy, University of Chicago and NBER. Email: jludwig@uchicago.edu. Sendhil Mullainathan: Department of Economics and Department of Electrical Engineering and Computer Science, MIT, and NBER. Email: sendhil@mit.edu.

1 Introduction

Education decisions increasingly depend on data, specifically achievement test data. These data are collected at great cost—schools now spend as much as 18% of all student time on testing or test preparation (DePaepe et al., 2015)—and inform a wide range of school decisions from which students to give extra help to which teachers to hire or fire. Test data have also become the lifeblood of education research, central to how we measure things like levels, trends, and inequality in educational performance, or the effects of policies like school spending, class sizes, teachers, teacher pay, charter schools, school choice, accountability systems, and much more.¹

While in one sense these test data are widely used, in another sense most of the data are thrown away. Each test contains dozens—sometimes hundreds—of individual items. But before those data ever reach any principal, teacher, parent or researcher, the item-level structure of the data is destroyed. What gets handed over instead is an aggregate measure like an average test score. Does deleting all of this item-level data wind up throwing away useful *information*?

The answer is not obvious *ex ante*. The leading theories about educational measurement would suggest that no useful information is lost. For example, the whole logic of item response theory models (IRT), which inform the design and reporting of these achievement tests, is that each individual test question captures a single underlying ability parameter (van der Linden, 2016). Is that view correct in practice? Or are we missing out on valuable information collected at great cost? No one knows, partly because the assumption that the aggregate measure is sufficient is so widespread and deeply ingrained that the item-level data are almost never even made available.

In this paper, we examine what, if anything, is lost by drawing on a unique dataset of item-level data from Texas. We have eight years' worth of Texas data on nearly 5 million K-12 students enrolled between 2012 and 2019. The data include 1.31 billion student-item responses together with rich information on each student including demographics, which teacher each student had, and outcomes like grade point average, school discipline, high school graduation, college attendance and even adult earnings. From these data we derive three key findings.

We first examine the assumption that conditional on the test average or IRT parameter, there is no extra information in the item-level data. We show this assumption is incorrect for both student and

¹While these are obviously vast bodies of literature, examples of influential research on the quality of the overall education system based on test data include, most famously, the Nation At Risk report (National Commission on Excellence in Education, 1983) and subsequent studies of international test data, e.g. Woessmann (2016). Influential examples of studies of test score inequality include Coleman (1966); Fryer Jr and Levitt (2004); Heckman and Krueger (2005); Jencks and Phillips (1998); Billings et al. (2014); Reardon (2018); Hashim et al. (2020). A vast econometric literature looks at the effects on test scores of policies such as school spending (Hanushek, 1986; Card and Payne, 2002; Jackson, 2020), class sizes (Hanushek, 1999; Angrist and Lavy, 1999; Krueger, 1999; Hoxby, 2000b), teachers (Kane and Staiger, 2008; Rothstein, 2010; Rockoff et al., 2011; Chetty et al., 2014b; Rothstein, 2017), teacher pay (Hanushek and Rivkin, 2006; Jacob, 2007; Barlevy and Neal, 2012), charter schools and school choice (Hoxby, 2000a; Cullen et al., 2006; Abdulkadiroğlu et al., 2011; Angrist et al., 2013; Dynarski et al., 2018; Walters, 2018), and accountability policies (Dee and Jacob, 2011; Figlio and Loeb, 2011).

teacher performance measurement. For example, we rank-order teachers on their “value-added” on an aggregate achievement measure, then compare that to teacher rankings of value-added on individual test items. If a single ability measure captures all the information in the test, then the rank-ordering of teachers by the aggregate versus individual items should be equal (up to statistical noise). Yet the average rank correlation of the aggregate and item-level orderings is just 0.66. For students, the corresponding average rank correlation is 0.75. We show that the deviation of these correlation estimates from 1 is not simply an artifact of noise in the item-level data.

Notice an important implication of these findings. In education there is an implicit assumption of “uniform dominance” that some students and teachers are uniformly better or worse than others. Our findings suggest that this is the wrong model of the world. A more accurate model seems to be “comparative advantage”: strong students and teachers still have areas where they can improve, while weaker students and teachers have areas of strength to build upon.

Second, we ask what that extra information is relevant for. We show that the items contain information about a wide range of decision-relevant student outcomes like class performance, suspensions, high school graduation, college attendance, or adult earnings. For example, we demonstrate that the item-level data help us better identify those teachers predicted to be in the bottom 5% of the value-added distribution in terms of predicted future outcomes for their students (Hanushek, 2009; Chetty et al., 2014b). We show that value-added measures based on aggregate data disagree with the (more powerful) item-level predictions 51.6% of the time for class failure, 42.4% for disciplinary infractions, 44.9% for high school graduation, and 39.6% for college attendance. In other words, an aggregator like the test score average or IRT parameter is not the optimal aggregator for these outcomes.

But the real problem here is not that we have the wrong aggregator, but rather that we aggregate at all. For example, if we take the optimal aggregator for high school graduation and the optimal aggregator for college attendance, they do not uniformly agree with one another. That is, not only do the individual test items contain useful “signal” for student outcomes, different items contain different signals for different outcomes. This has implications for the economics literature on “anchoring,” which solves the problem of comparing learning gains at different ages by linking or anchoring test scores with some future outcome (Bond and Lang, 2018; Cunha and Heckman, 2008; Cunha et al., 2010; Nielsen, 2019). Our results imply the choice of specific outcome on which to anchor is not innocuous. More importantly still, our findings also imply that education practice cannot avoid wrestling with normative judgments about education’s goals.

Third, we show that there are practical benefits from working with item-level data. Many education decisions hinge on predictions of future student outcomes, and so are examples of “prediction policy problems” (Kleinberg et al., 2015, 2018; Rambachan, 2024). That means anything that improves the prediction of future outcomes can improve decision quality. Framed that way, no one could possibly argue that the best way to predict future outcomes is to rely on the aggregate test score and throw away all the information from the high-dimensional item-level data.

Can these predictions actually be used in practice? Notice a key challenge: Items on a given test change year-to-year. So a predictor formed on historical test data can't be applied to future years with new questions. Solving this problem requires connecting questions from past exams to new unseen questions. While that could best be done with natural language processing, we demonstrate here that even categorizations that do not use text can already yield value. To link “like” questions across years we offer a “proof-of-concept” implementation that creates item categories based on how the categories differentiate teachers with respect to their patterns of comparative advantage. We do not claim this is the optimal categorization; to the extent to which this is not optimal, we understate the gains from using item-level data. We show that, relative to identifying the bottom 5% of teachers in the distribution of graduation value-added based on aggregate scores, using item-categories increases the impact of the policy on student outcomes by 21%.

What is particularly striking about these potential gains is the low cost at which they can be derived. The marginal costs of storing, transmitting, and analyzing data at the item level rather than the aggregate level are trivial, both absolutely and certainly as a share of education spending. We show that the gain from switching the basis of education decisions from aggregate test scores to test items yields an infinite marginal value of public funds (Hendren and Sprung-Keyser, 2020) even under quite pessimistic assumptions about cost.

The results we present here demonstrate empirically that the current way we use test information gives us a distorted view of what is happening. We are throwing away useful information that could help us both create a better educational system and do better research on that system. We need, in other words, a new paradigm for using test information.

2 Related Literature

Our paper builds on the large psychometric literature on educational testing and how to interpret the results of these tests.² With the rise of standardized testing in the early 20th century, researchers were confronted with a variety of challenges relating to how to interpret their results. For example, how should researchers equate student performance across tests comprised of different test items? Item response theory (IRT) helped solve this problem by viewing each item as a window into some single latent student ability, where test items may differ in how much information they convey about that underlying ability.

The assumption of a single latent student ability has since been relaxed by multidimensional item response theory (MIRT) (Reckase, 1985; van der Linden, 2016). Once multiple factors are allowed, however, the natural question becomes how to group items into factors. MIRT models form these groupings statistically, aiming to explain variation across test items. But the goal of education is

²For overviews see, for example, Hambleton et al. (1991), van der Linden and Hambleton (1997), van der Linden (2016), and de Ayala (2009). For recent psychometric work connecting test items to treatment effect heterogeneity, see Gilbert et al. (2024, 2025) and Ahmed et al. (2024).

not to explain variation across test items, it is to improve student learning and life outcomes. That requires making decisions that affect these outcomes. The factors that best explain item variation (the focus of MIRT) need not be those that are most relevant for these decision-relevant outcomes. This may help explain MIRT's limited use in practice.

Economists have thought about connecting test scores to student outcomes in the context of solving a different specific problem: the concept that test scores aim to measure, “knowledge,” has no natural scale (unlike earnings, mortality rates, widget production, etc.). Consequently, test scores are typically reported on ordinal scales (i.e., rank-ordering). The only way to interpret a “unit” difference in test scores is to adopt some measurement model. Research shows that key conclusions about these magnitude questions—like whether the Black–White test score gap widens or narrows as students move through school, the extent of dynamic complementarity, or whether intervention effects persist or fade—can be highly sensitive to the choice of measurement model and scaling method (Bond and Lang, 2013; Cascio and Staiger, 2012; Nielsen, 2019; Agostinelli and Wiswall, 2025).

In response, economists have “anchored” test scores to some economically relevant future outcome; that is, anchoring rescales test scores based on their predictive relationship to the outcome. This allows researchers to interpret test score intervals with respect to differences in the anchoring measure, such as educational attainment (Bond and Lang, 2018) or earnings (Cunha and Heckman, 2008; Cunha et al., 2010). But this anchoring approach still assumes that students with the same score possess the same ability—even if they reached that score through different item-response patterns. That is, the anchoring approach still assumes that a single summary measure of student performance captures all the relevant information contained in the test.

A few previous papers have implicitly or explicitly questioned the idea that standardized tests capture only a single latent student ability. For instance, Bettinger et al. (2013) find that scores on the Math and English components of the ACT better predict college success than do scores on the Reading and Science components. The implication is that predicting college success using an overall ACT score would throw away information compared to relying on subject-specific ACT results, although our results suggest that even relying on subject-specific tests throws away a great deal of information contained in the items within each test. Ding et al. (2023) investigate the gender gap in science and show that even the sign of the gap depends on the domain of scientific intelligence being assessed: girls outperform boys at identifying scientific issues but under-perform in incorporating existing knowledge in new situations.

Most similar to our own work is Nielsen (2019), who, like us, also notes that the test items that get the most weight in psychometric models need not be those that are most relevant for decision-relevant criteria such as medium- or long-term student outcomes. The primary focus of his analysis is to show how weighting items differently according to their relationship to different long-term outcomes yields even larger estimated achievement gaps by race, gender, and income. Our study complements his in that our main focus is instead on formalizing and testing the assumptions be-

hind the current approach to test-score aggregation, and on empirically demonstrating that feasible approaches to making greater use of the item-level data can improve concrete educational decisions.

Within education, our work also adds to a recent literature that applies machine learning methods to non-traditional (often high-dimensional) data to measure complex educational concepts.³ We are aware of four papers in this literature that, like ours, apply modern machine learning methods to high dimensional data sources. [Adukia et al. \(2023\)](#) applies machine vision to children’s books to measure stereotypes, while [Adukia and Harrison \(2025\)](#) examine the similarities and differences in curricula across state public school systems and between public and private schools. [Biasi and Ma \(2022\)](#) apply natural language processing (NLP) techniques to course syllabi to measure frontier learning in higher education. [Lee and Schaelling \(2024\)](#) use NLP methods to measure bias in reading passages embedded in standardized test questions.

Finally, our study relates to a growing body of work on so-called “prediction policy problems” ([Kleinberg et al., 2015](#); [Mullainathan, 2025](#)), which focus on decisions that hinge on a prediction of some outcome. For example, judges decide whether to release people awaiting trial based on predicted flight or re-arrest risk ([Kleinberg et al., 2018](#); [Rambachan, 2024](#)), HR managers hire based on predicted worker productivity ([Li et al., 2025](#)), guidance counselors recommend college courses based on predicted student performance ([Bergman et al., 2021](#)), social workers investigate child abuse claims based on predicted allegation veracity ([Chouldechova et al., 2018](#); [Grimon and Mills, 2025](#)), tax authorities audit returns based on predicted fraud ([Battaglini et al., 2025](#)), while lenders give credit based on predicted default risk ([Blattner and Nelson, 2021](#); [Rambachan et al., 2022](#)). We argue that education decisions frequently share this same structure, which requires changing how we think about the use of testing data.

3 Conceptual Framework

Schools spend a great deal of time both deciding what information to collect and then actually collecting it. What students should know has been a flashpoint in education policy debates for decades: the so-called “curriculum wars” ([Loveless, 2014](#)). Schools then devote vast resources to measuring student progress toward those goals; in some districts, as much as one of every five student hours in school is used for testing or test prep ([DePaepe et al., 2015](#)). All that information is then collapsed into a single number. What does that aggregation assume, and what might be missed?

Let d_{iq} indicate whether student i answered question q correctly. Let the vector of student responses

³There is also a literature that uses educational data besides test scores to measure different educationally relevant concepts. For example, [Brown et al. \(2025\)](#) and [Jackson et al. \(2020\)](#) use survey instruments to measure concepts such as cognitive endurance and socioemotional development. Similarly, [Jackson \(2018\)](#) uses non-traditional administrative data such as suspensions and grade progression to develop novel measures of non-cognitive skill.

on a given test (T) be $D_i = \{d_{iq}\}_{q \in T}$.⁴ However, that vector of individual item-level responses is rarely provided to key decision-makers such as teachers, school administrators, parents, and students. What these stakeholders typically get instead is the output of some aggregator; see, for example, Figure 1, which shows a sample report for the state from which our own dataset is drawn (Texas). We define an aggregator as a mapping from student item-level responses to a scalar:

$$a_i = a(D_i) \tag{1}$$

[Figure 1 about here.]

Common aggregators used for educational policy include:

- The *simple averaging* aggregator: $a(D_i) = \bar{D}_i = |T|^{-1} \sum_{q \in T} d_{iq}$.
- The *percentile rank* aggregator: $a(D_i) = F(\bar{D}_i)$ and where $\bar{D}_i \sim F$.
- The *three parameter IRT* aggregator: $a(D_i) = \arg \min_{\theta} - \prod_{q \in T} P_q(\theta)^{d_{iq}} (1 - P_q(\theta))^{1-d_{iq}}$, where $P(d_q = 1|\theta) = P_q(\theta) = c_q + \frac{1-c_q}{1+e^{-d_q(\theta-b_q)}}$, so that θ denotes the uni-dimensional latent ability of the student, and (c_q, b_q) are question-level psychometric parameters typically identified during question development and testing.⁵

The key feature of these approaches is the assumption that all relevant information in the vector of item-level data about how students, teachers, or schools are doing can be captured by a single scalar. The other key feature of these aggregators is that the choice of that scalar is based purely on *statistical* considerations. For example, IRT models are designed with the statistical goal of explaining as much variation in the item-level data as possible (van der Linden, 2016).

But the only people who care intrinsically about statistics are statisticians. Everyone else—educators, parents, students, and education economists—care about *decisions*. Which students should pass a class and advance to the next grade or receive a high school diploma?⁶ Which students should be prioritized for academic remediation efforts like summer school (Jacob and Lefgren, 2004) or high-dosage tutoring (Guryan et al., 2023; Bhatt et al., 2024; Nickow et al., 2024)? Which teachers should be paid more, retained, or promoted (Kane and Staiger, 2008; Chetty et al., 2014b; Hanushek et al., 2023; Bleiberg et al., 2025)? Which schools are worth attending, or should be put under new management (Kane and Staiger, 2002; Abdulkadiroğlu et al., 2011; Dem-

⁴This definition is without loss of generality since we can redefine the notion of a question to encompass other relevant features of the items. For example, if we want to leverage information about what specific “incorrect” answer a student gave, we could define binary variables with a question index at the item-by-response level (e.g., 1A, 1B, 1C, 1D, 2A, 2B, . . .). Similarly, we could add a time dimension to the model by redefining the question index to be a question number-by-year combination.

⁵As noted above, there is a small literature on multiple IRT (MIRT) models in education, which assume multiple abilities rather than a single ability; however, these models are rarely, if ever, used in practice by school systems in the US.

⁶For examples of policies in Texas that condition grade promotion or diploma receipt on some achievement measure, see: <https://tea.texas.gov/student-assessment/testing/student-assessment-overview/texas-assessment-program-faqs-04.04.18.pdf>.

ing, 2014; Angrist et al., 2017)? Which high school graduates should be admitted to which college (Hoxby and Avery, 2012; Arcidiacono and Lovenheim, 2016)?

Note that these decisions share a common structure: Someone confronts some choice variable X , made on the basis of some observable feature Y , chosen to maximize some objective function $\pi = \pi(X, Y)$. Each decision is thus represented by the triplet:

$$\mathbb{D} = (\pi, Y, X). \quad (2)$$

A major challenge in maximizing welfare is that, in practice, Y is either unknown when the decision is made or only imperfectly observed. What trajectory is a student on in terms of their learning or life more generally (high school graduation, college, future livelihood)? How effective is a given teacher or school? In those cases, the decision-maker’s task reduces to a “prediction problem” (Kleinberg et al., 2015): What is the best way to use currently observable student data, such as test performance (D), to predict the outcome of interest (Y) that determines the welfare impact of the decision (X)? These decisions involve assigning students, teachers, schools, or even entire school systems to some intervention based on some rank-ordering of the expected Y value, with $\frac{\partial \pi}{\partial Y} \frac{\partial Y}{\partial X} = 0$, distinct from decisions that hinge on a causal rather than predictive inference, where $\frac{\partial \pi}{\partial Y} \frac{\partial Y}{\partial X} \neq 0$.

We can now state more precisely our motivating question: what costs arise from the hidden aggregation of information? By this we mean: What happens to decision quality when we base some decision-relevant prediction not on the full vector of each student’s item-by-item test performance, D , but rather on the output of some aggregator, a ?

Call an aggregator $a^*(\cdot)$ *optimal* for decision \mathbb{D} if predictions of Y using a^* minimize welfare loss relative to the ideal policy chosen when Y is perfectly observed or predicted.⁷ For simplicity, we restrict our attention to the class of welfare functions that imply mean squared loss with respect to prediction errors.

What assumption does aggregating student achievement make, even if this assumption is usually made only implicitly? Suppose that student i ’s performance on question q follows a single threshold-crossing model:

$$d_{iq} = \mathbb{1}(\beta_q \theta_i > \epsilon_{iq}) \quad (3)$$

where $\theta_i = \{\theta_{is}\}$ is a vector of latent skills s , $\beta_q = \{\beta_{qs}\}$ describes the sensitivity of question q to skill s , and $\epsilon_{iq} \sim \mathbb{U}(0, 1)$ is normalized to be a uniformly distributed performance error. Clearly, when $\dim(\theta) = 1$, so that learning is uni-dimensional, it is possible to aggregate without losing

⁷Note that, because test questions are not repeated year-over-year, some form of aggregation is unavoidable. Otherwise, there is no way to apply relationships learned from data where Y is available (e.g., cohorts of students who have already graduated high school) in order to make decisions for samples where it is unknown or imperfectly observed (e.g., cohorts of students who have not yet graduated from high school). We return to this point in Section 7.

consequential information about student achievement. That is, predicting Y using a^* should be no worse than predicting Y with the item-level vector of student responses, D . But what happens if θ is *multi-dimensional*? The following proposition clarifies the conditions under which it is sensible to aggregate test score data into a single measure of student achievement intended to serve as an input for multiple decisions.

Proposition 1: The following statements are equivalent.

1. **Single latent factor model**: $\forall s \neq s'$ and $\forall q$ we have that $\beta_{qs} = 0$.
2. **Single optimal aggregator**: $\exists a^*$ that is optimal for every \mathbb{D} .

Proof: See Appendix B.

Intuitively, Proposition 1 says that the decision by policymakers to rely on a single aggregation of the item-level data is, in the best-case scenario, equivalent to assuming that the systematic component of student test performance is uni-dimensional. Therefore, if the existing aggregation method is optimal for all of the real-world decisions in which it currently serves as a critical input, then all remaining variation should just be noise around the corresponding conditional mean. This proposition is illustrated intuitively in Figure 2.

[Figure 2 about here.]

Whether the existing aggregator is optimal for every decision, or whether an aggregator that is optimal for every decision exists *at all*, is ultimately an empirical question. In the sections that follow, we explore the implications of relying on a single aggregator for understanding student achievement and teacher value-added, and the potential downstream effects for decisions that use test score data as a critical input.

4 Data and summary statistics

We draw on administrative data from the Texas Education Agency (TEA) that cover all public-school students enrolled in the state of Texas for the time period 2011-12 through 2018-19. The dataset includes detailed demographic information about students, including student demographics, test score data that includes item-level responses, medium- and longer-term student outcomes, and information about which students had the same teacher.

The student demographic information includes ethnicity (six ethnic groups), economically disadvantaged status (four groups),⁸ gender, limited English status (three groups),⁹ special education

⁸These groups are: not disadvantaged, free lunch eligible, reduced-price lunch eligible, and ‘other economic disadvantaged,’ which are students determined to be economically disadvantaged through means other than the National School Lunch Program, such as: being from a family with income below the federal poverty line, Temporary Assistance to Needy Families eligibility, or Food Stamps eligibility.

⁹These groups are: current, former, and never.

status, and gifted status. Demographic coverage is nearly universal, with missing data for less than 0.1 percent of observations. (We use missing data categories to deal with the limited missing demographic data.)

The data allow student-teacher linkages, but do not include either teacher identifying information or much information about the characteristics of these teachers. Nonetheless, knowing which students had the same teacher for the same subject in the same year lets us calculate teacher value-added estimates of the sort that are increasingly used in education policy decisions and education research (Kane and Staiger, 2008; Chetty et al., 2014a; Koedel and Rockoff, 2015).

The dataset also includes results for the State of Texas Assessments of Academic Readiness (STAAR), which are designed to “measure the extent to which a student has learned and is able to apply the defined knowledge and skills in the Texas Essential Knowledge and Skills (TEKS) at each tested grade, subject, and course.”¹⁰ We focus on the STAAR tests for math and reading for students enrolled in grades 3-8, so our sample consists of students who are enrolled in one of those grades for at least part of our study period. We focus on the standard version of the state test,¹¹ a multiple-choice test with four response options for each test item.

- Math tests typically contain 46-56 items (reduced to 32-42 items starting in 2016-17).
- Reading tests contain 42-52 items (reduced to 34-44 starting in 2016-17).¹²

Our item-level data record the exact answer the student gave to each question (e.g., answer “B” was given for question 1) or if the student did not answer the question.¹³ Our analysis period starts in 2011-12 because that is the first year the item-level test data are available.

Our data can also be matched to several relevant medium- and long-run outcomes such as classroom academic performance (e.g. class failure), disciplinary infractions, high school graduation, college attendance, and, for a single cohort of 8th graders, earnings. High-school graduation data are recorded directly by the Texas Education Agency (TEA), while college records come from linking students in the TEA data to administrative records from the National Student Clearinghouse, which captures enrollment at the near-universe of colleges in the country.¹⁴ These long-run data only cover earlier cohorts, since later cohorts have yet to reach the required age to achieve that outcome (e.g., third grade students in 2018-19 have not yet graduated). As our long-run outcome data

¹⁰<https://tea.texas.gov/student-assessment/staar>.

¹¹We therefore exclude alternative versions of the STAAR including STAAR Spanish, STAAR Modified, and STAAR Alternate which are designed for English Learners or students receiving special education services.

¹²The question range in both the math and reading tests is due to the fact two additional test items are added each grade. Therefore in 2016-17 the third-grade math test had 32 items, the fourth-grade math test had 34 items, etc.

¹³Except years 2011-12 (math and reading) and 2014-15 (math only) where the data only record whether the student provided the correct answer or did not answer the question.

¹⁴Unfortunately, our data from the National Student Clearinghouse only cover up to 2019. We therefore use data from the Texas Higher Education Coordinating Board to obtain college enrollment outcomes up to 2021. The drawback of these data are that they only contain in-state college enrollment. Fortunately, we expect the impact of this data limitation to be negligible since Texas has fairly limited out migration of college students in practice: only 3.7 percent of students from Texas enroll in an out-of-state college (Mountjoy, 2022).

cover up to 2021, we restrict high school graduation to students entering third grade in 2011-12 or before, and college-going to students entering fourth grade in 2011-12 or before.¹⁵

For our first cohort of 8th graders, we can also match these students to earnings data from the state unemployment insurance system.¹⁶ This captures earnings in UI-covered jobs 5 years after high-school graduation.¹⁷

Table 1 provides summary statistics for our dataset. Column (1) does so for all students enrolled in grades 3-8 during our study period (4.91 million total), while column (2) restricts to the 4.65 million test-takers with valid test scores.¹⁸ We are able to link a total of 3.65 million students to information about what teacher they had along with the lagged test score necessary to calculate a teacher value-added measure; this sample is shown in column (3).¹⁹

Texas public schools are majority Hispanic (51.8%), with white and Black students making up 29.5% and 12.7% of the student body, respectively. 60.7% of students are categorized as economically disadvantaged. Comparing columns (1) with (2) and (3), we see that the overall student population is generally similar to our two analysis samples. Our samples are slightly less likely to be limited English proficient or receive special education services, because a substantial portion of both groups take alternative assessments rather than the tests we examine here.

The top panel shows data for the 1.31 billion student-item-level observations we have; students on average get 61.6% of math items correct and 66.5% of items right for English Language Arts (ELA) (column 2). The figures are similar for the sub-sample we can link to teachers.

[Table 1 about here.]

¹⁵For example, if third-grade cohorts from 2011-12 or before are covered, then the linkage will include: third-grade students in 2011-12 only, fourth-grade students up to 2012-13, fifth-grade students up to 2013-14, sixth grade students up to 2014-15, seventh-grade students up to 2015-16, and eighth-grade students up to 2016-17.

¹⁶Earnings records are linked to the education data by the Texas Education Agency. Given the overall low levels of out-migration in Texas, we expect this variable to capture labor force outcomes for the majority of our sample (Mountjoy, 2022).

¹⁷Around 90% of US workers are in jobs covered by UI; see fredblog.stlouisfed.org/2023/12/the-insured-unemployment-rate. For a discussion of which jobs are eligible for UI coverage see: oui.doleta.gov/unemploy/pdf/uilawcompar/2023/complete.pdf.

¹⁸The test-taking sample requires valid tests in both math and reading. We define a valid test as a standard STAAR test that does not contain missing answers for all questions.

¹⁹Our value-added sample requires valid lagged test scores, which necessitates: (i) dropping third grade from the sample (roughly 2.5 million student-year observations), and dropping an additional 0.5 million student-year observations lacking valid lagged test scores. We also must be able to uniquely match students to their teachers, eliminating 2.7 million student-year observations for which we cannot create a unique match. We match students to both their math and reading teachers, so this match varies slightly by subject. Column (3) of Table 1 reports the value-added sample for the math sample. We also drop roughly thirty-thousand student-year observations where the class size is below five or above two-hundred.

5 Is academic “ability” really uni-dimensional?

Does hidden aggregation destroy information in a way that degrades decision quality? Our first step is to ask: Are the data consistent with the idea that a single parameter can adequately capture all the information that exists in the item-level test data? The answer is no.

We start off by examining student performance on different sub-categories of test items. The “single ability” model implies that students who are particularly good on one sub-category of test item should be particularly good on every other sub-category as well, since they all reflect the same underlying latent construct. But that is not what we see; students have different strengths and weaknesses. We then show that the same thing is true for teachers with respect to their value-added to student learning. In fact, aggregation winds up destroying *most* of the predictable variation connected to teacher performance.

5.1 Student performance

The single latent factor model implies that any variation that remains after conditioning on “the” measure of ability (average performance, θ , etc.) must be noise. To see why, observe that equation 3 implies that we can write the probability of correctly answering item q as:

$$p_{iq} = \beta_q \theta_i \tag{4}$$

Equation 4 shows that, if the single-latent-factor model holds, then there are only two systematic sources of variation in the item-level data: uni-dimensional student ability (θ_i) and question difficulty (β_q). In expectation, student performance will therefore be identical for all test items after conditioning on question difficulty.

This is a testable prediction. Intuitively, we would like to regress student-by-item-level response data on question fixed-effects, to control for item difficulty, and student-by-item fixed effects, to capture the possibility that student ability differs across different test items. Estimating the equation in this exact form is infeasible, however, because students answer each item only once on a given test, so student-by-item interaction terms are not identified.

However, we *can* test a closely-related implication of the single-latent-factor model: if expected student performance is identical on each and every item after conditioning on item difficulty, it follows immediately that student i ’s performance should also be identical on every *subset* of the test after conditioning on item difficulty. We explore this implication empirically by estimating fixed-effect regressions of the following form:

$$d_{iq} = \omega_q + \gamma_{ic(q)} + u_{iq} \tag{5}$$

Where ω_q is a question fixed effect; $c(q)$ is a function that maps questions into mutually exclusive

subsets of items that we will call categories; and $\gamma_{ic(q)}$ is a student-by-category fixed effect. The parameters of interest from model 5 are the student-by-category interactions ($\gamma_{ic(q)}$). The question fixed effect (ω_q) will fully account for variation in performance due to question difficulty (see equation 4). Our preferred model uses a k-means clustering algorithm to create a data-driven categorization of the items (Hastie et al., 2009); however, we find similar results using the categorizations provided by the test maker, or under an alternative categorization that uses patterns across teachers in their comparative advantage in teaching different types of material (see appendix A.1 for more detail on the clustering algorithm and these various robustness checks).

We view these estimates as likely reflecting a lower bound on the hidden structure for students. Our student-by-category fixed effects will still miss systematic variation if there is: (1) a one-to-many mapping between items and ability; (2) more dimensions of student ability than categories;²⁰ and/or (3) an imperfect correspondence between our categories and the actual dimensions of learning. However, to the extent that these sources of mis-specification are a problem, they work in the direction of leading us to *understate* how much additional information there is in the disaggregated test-score data relative to the usual aggregate measures (test average, theta, etc.).

Figure 3 reveals systematic variation in the student-by-category fixed effects, by showing how students swap ranks in the distribution of performance across test-item categories. The column furthest to the left shows a rank-ordering of students created using their average performance across all questions.²¹ The remaining columns along the X-axis represent performance *within* each test item category (that is, alternative rankings created using the student-by-category fixed effects). The Y-axis denotes a student’s rank in the distribution of $\gamma_{ic(q)}$ within that category. To visualize the degree to which students are ranked similarly for each test category as for the overall test average, we color-code each student based on their rank in the overall test-score distribution (from the left-most column), and then for each test-item category we rank them in that specific distribution but preserve their overall test-rank color-coding.²²

If a single latent ability best explained student performance on academic achievement tests, we would expect to see the same rank-ordering of students on each sub-category of the test as we observe using the overall average test performance (up to noise in the data). Visually, this would result in each vertical bar in figure 3 having an identical color gradient, since each color-coded rank-ordering would be the same for the overall average test performance and performance on each sub-test category.

²⁰We chose the number of categories for our data driven method to align with the number of official content categories provided by the state of Texas; however, we find similar results if we use the state content categories or alternative clustering algorithms (see appendix figure A.1).

²¹We measure this by estimating model 5 while imposing that student effects are homogenous across items: $\gamma_i = \gamma_{ic(q)} \forall q$, which is equivalent to de-meaned average performance: $T^{-1} \sum_T d_{iq} - \bar{d}$.

²²So for example a student who scores well on the test average will be coded as blue and plotted towards the top of the left-most column in the figure; if they perform badly in the first sub-category of test items shown in the second column, they will be plotted towards the bottom of that column but will stay color-coded as blue (reflecting their average test score performance).

Instead, we see a great deal of “rank-swapping,” denoted visually by a mixing of colors across columns. That is, the students who are relatively strongest on average (those ranked at the top of the graph according to the left-most column for average performance) are relatively weak on a number of test sub-categories, while students who are relatively weak on average (ranked at the bottom of the left-most column for average performance) often have sub-categories where they show signs of strength. The results in the figure are for a 1% sample of randomly selected students that were in grade 4 in year 2016; however, we find similar patterns for other grades and years (see table 2 for a comprehensive summary of information loss in each grade and year).

[Figure 3 about here.]

Taking the results across all of the test sub-categories, the average rank correlation of student performance on a given test sub-category with the overall average test performance is 0.75 (min = 0.34, max = 0.92 – see Appendix Figure A.5 for the full distribution). These results suggest there is additional structure in the items that is hidden by aggregating.

This variability in ranking across categories is not simply due to sampling variation. To show this, we carry out a placebo study. Panel (b) of figure 3 presents results from a simulation of the expected amount of rank-swapping under the single-latent-factor null, with the distributions of noise and student performance calibrated to match the relevant moments of our data.²³ We find that some rank-swapping is expected purely as a result of sampling variation; however, the visual benchmark in panel (b) exhibits substantially less mixing than the actual data, which suggests that chance alone cannot account for these results.

These results leave unanswered whether this hidden structure represents a large versus small share of the overall information about student performance that is contained in the test. Exactly how much systematic variation do we lose by ignoring the item-level structure? Estimating the amount of “information loss” also lets us formally calculate a statistical test for the degree to which noise explains our results.

Observe that we can define the total amount of predictable variation (up to the need for categorization) in the test as $var(\gamma_{ic})$. Aggregation destroys information by eliminating the category-by-category variation around some estimated mean student effect. Thus, the “predictable variation lost” is given by: $var(\gamma_{ic} - \gamma_i)$, where $\gamma_i = |T|^{-1} \sum_{q \in T} \gamma_{ic(q)}$ or, equivalently, the coefficient from model (5) after imposing the appropriate linear restriction ($\gamma_{ic} = \gamma_i \forall q$). Thus, we quantify information loss as the share of all predictable variation that is missing due to implicitly (or explicitly) assuming ability is uni-dimensional when aggregating:

$$\text{Information loss (\%)} = \frac{\text{Predictable variation around the mean}}{\text{Total predictable variation}} = \frac{var(\gamma_{ic} - \gamma_i)}{var(\gamma_{ic})}. \quad (6)$$

²³To be precise, we assume each student’s performance for a given category is equal to their average score, and then we add noise to each category score where the noise distribution is assumed to be normal and with variance equal to the level of noise implied by the category-level standard errors. See appendix A.2

In practice, we estimate the numerator and denominator necessary to construct the information loss metric using a simple variance decomposition of the raw item data (see appendix A.3 for more detail).

To account for sampling variation, we also report F-statistics and P-values from a formal test of the null that there is no information loss (equivalently, that there is a single latent factor). To see how we construct the test statistic, observe that estimating the student-level average score is equivalent to imposing a linear restriction ($\gamma_{ic} = \gamma_i \forall c$) on equation (5). Thus, testing a null of no information loss is equivalent to a simple test of nested linear models (Wooldridge, 2010).²⁴

The results of these calculations for students are contained in Table 2. Each row denotes a separate grade of enrollment. The first four columns denote AY 2016-2019, while the fifth column (labeled 16-19) summarizes average information loss over all years within a given grade.

[Table 2 about here.]

We find that an average of 14.8% (ranging from 9-20% depending on the particular year and grade level) of the available information about students is lost. The p-values from the joint F-test confirm that these results are inconsistent with finite sample noise or over-fitting as sufficient explanations for the patterns we see in the data.

Note that we directly account for the potential of over-fitting in finite samples with the high-dimensional fixed-effect model (equation 5) used in the variance decomposition that generates the information loss metric. First, we note that the estimates of mean squared error used in the variance decomposition to calculate information loss apply the standard degrees-of-freedom correction, a procedure that yields unbiased estimates of residual variance (see appendix A.3). Second, we note that the test statistic we deploy to directly test the null of no information loss is not sensitive to the number of model parameters being estimated (Wooldridge, 2010).

These point estimates are almost surely a lower-bound estimate for the share of student information destroyed. The reason (as discussed above) is that for our analysis of student performance we are forced to rely on an imperfect categorization of test items, rather than being able to directly estimate the probability each student has of correctly answering each item. Thus, we will mechanically understate the magnitude of the loss if, for example, there are questions that test multiple skills. Consistent with this hypothesis, we will show next that for teachers, where we can identify actual teacher-by-item fixed effects, we see much greater information loss.

5.2 Teacher performance

We showed above that for the analysis of student test data, the single-latent-factor model implies that any variation that remains after conditioning on “the” measure of ability (average performance,

²⁴ $F = \frac{(SSR_r - SSR_u)/m}{SSR_u/(n-k-1)}$ where SSR_r and SSR_u are the restricted and unrestricted sum of square errors, m is the number of restrictions, k is the number of parameters, and n is the number of observations.

θ , etc.) must be noise. For teachers, equation 4 has a similar implication.

To see why, decompose student ability as: $\theta_i = \theta_{i_0} + \delta_{j(i)} + \eta_i$, where θ_{i_0} is the student’s ability at baseline; $j(i)$ is an index that denotes the teacher $j = j(i)$ assigned to student i for the academic year preceding the test; $\delta_{j(i)}$ is the component of ability growth explained by teacher j ; and η_i is mean zero, idiosyncratic growth in student ability. Thus, under the null of the single-latent-factor model, we can write down a corresponding decomposition of the systematic component of student performance on item q as:

$$p_{iq} = \beta_q(\theta_{i_0} + \delta_{j(i)} + \eta_i) = p_{qi_0} + p_{qj(i)} + u_{qi} \quad (7)$$

where $p_{qi_0} = \beta_q\theta_{i_0}$ is the likelihood that student i would answer question q correctly at baseline; $p_{qj(i)} = \beta_q\delta_{j(i)}$ is the component of growth in performance on item q explained by teacher j ; and u_{iq} is residual growth in the likelihood of answering question q correctly that is not otherwise predicted by teacher assignment.

Equation 7 reveals that, as with students, the single-latent-factor model of decision making implies that the component of growth predicted by teachers ($\beta_q\delta_{j(i)}$) should be *identical* (in expectation) for each and every item after conditioning on question difficulty. For teachers we can test this implication empirically by estimating teacher value-added-style fixed effect models analogous to equation 7:

$$d_{iq} = \omega_q + \delta_{j(i)q} + \pi X_i + u_{iq} \quad (8)$$

where ω_q is a question fixed effect; $\delta_{j(i)q}$ is a teacher-by-question fixed effect; and X_i are student-level controls. The parameters of interest are the teacher-by-item interaction terms ($\delta_{j(i)q}$). The vector of controls (X_i) will always contain an average score from the prior year’s test and hence account for baseline student ability.²⁵ In addition, our baseline model also includes the “standard” vector of teacher value-add controls that have been shown to generate unbiased or nearly unbiased estimates of the causal effect of teacher j on test scores (Chetty et al., 2014a; Koedel and Rockoff, 2015; Bacher-Hicks et al., 2019).²⁶ At the same time, we note that our test of the single-latent-factor model does *not* require that our teacher value-added estimates are causal. Our view is that these teacher value-added models are so widely used for education decision-making that it is important to explore the information that might be missed by hidden aggregation regardless of the claim of the value-added approach to causality.

²⁵Observe that, under the single-latent-factor model, the average of last year’s items is simply $\bar{\beta}\theta_{i_0}$ and hence, up to a normalization, is equivalent to baseline student ability. To conform with prior literature on teacher value-added, our estimating equation includes lagged average scores, although our results are robust to including question-by-lagged score interactions as would be implied by equation 7 (see appendix A.4 for more detail).

²⁶See appendix A.5 for the precise list of controls.

We account for the classroom-shock component of teacher value-added variance by adapting the jackknife empirical-Bayes procedure from [Chetty et al. \(2014a\)](#) to the item level (see appendix [A.6](#)). However, we find similar results when directly examining the fixed effects from equation [8](#) (see appendix [A.4](#)). As is typical when estimating teacher value-add, we test our estimates for forecast bias. We find that our procedure yields item-level teacher value-add estimates that are nearly forecast unbiased for teacher-by-year item-level growth.²⁷ See appendix [A.6](#) for details regarding implementation of the empirical-Bayes procedure.

Figure [4](#) has the same structure as our previous figure for students, where we rank-order teachers based on their value-added on average test scores (which is 4th grade math in the figure), and assign them a color for that ranking (first column), then show where those teachers fall in the ranking for each individual math test item in subsequent columns — that is, teacher rank on the relevant teacher-by-question interaction ($\delta_{j(i)q}$).²⁸ We plot results for all 2,280 teachers that taught grade 4 in year 2016 and that met our estimation criteria; however, as with the student level analysis from the preceding section, we find similar patterns for other grades and years (see [table 3](#) for a comprehensive summary of the patterns of information loss).

[Figure 4 about here.]

Visually, we can see signs of even *more* rank-swapping across test components for teacher performance compared to what we saw for student performance in the previous sub-section. Across all items, the rank correlation between teacher value-added on the average score versus the item level measure averages 0.66 (min = 0.17, max = 0.86 – see Appendix Figure [A.5](#) for the full distribution), once again suggesting that the average score is insufficient for characterizing *any* of the underlying patterns that determine average teacher performance. These results suggest that there are teachers at the bottom of the overall distribution that exhibit areas of real strength that could be cultivated and leveraged. There are also teachers with excellent overall value-added who exhibit areas of real weakness that could, potentially, be improved if teachers were provided constructive feedback in their deficient areas. However, all of that information is lost when we focus on the average alone.

As with the student results, we address the potential for noise to generate rank-swapping with a placebo study. Panel (b) shows the amount of rank swapping we would expect to see due to noise alone.²⁹ The visualization strongly suggests that the results in Panel (a) are not simply a product

²⁷Our item level forecast-bias coefficient is 0.939. While this statistically differs from one – which would indicate forecast unbiasedness – we suspect that this is driven by the fact we only allow one year of drift since only three years of data are used to estimate teacher-by-item VA for computational tractability (see appendix [A.6](#)). Indeed, our estimate lies between models with and without drift: [Chetty et al. \(2014a\)](#) allows for seven years of drift and finds a forecast coefficient of 0.998, while a prior version which did not incorporate drift obtained a forecast coefficient of 0.861 ([Chetty et al., 2011](#)).

²⁸We measure teacher value-add on average scores by estimating model [8](#) while imposing that the component of growth predicted by teachers is homogeneous across items: $\delta_j(i) = \delta_{j(i)q} \forall q$. Up to a normalization, this is numerically equivalent to estimating teacher value-add on average scores (see appendix [A.7](#) for more mathematical detail).

²⁹More precisely, panel (b) of figure [4](#) presents results from a simulation of the expected amount of mixing under the

of sampling variability. (As with our results for students, we formalize this with a statistical test when computing summary measures of “information loss.”)³⁰

Table 3 presents results that quantify the share of predictable variation lost for each year and grade in our data. These results are analogous to those presented for students in table 2.³¹ As with the results for students, we account for sampling variation by reporting F-statistics and P-values from a formal test of the null that there is no information loss (equivalently, that there is a single latent factor).³² We note here that we perform the relevant residual variance decompositions necessary to generate the estimates in table 3 using jack-knife, leave-year-out estimates of item-level teacher value-add (see appendix A.6). Thus, for teachers, the estimates in this table are effectively generated out of sample and therefore not subject to over-fitting.

[Table 3 about here.]

For teachers, we find that a *majority* of the available information about teacher performance is lost as a result of aggregation, with the point estimates suggesting that the scale of information loss averages 65.6% (ranging from 58-74% depending on the particular year and grade level). This suggests that the decision to focus on average scores for the purpose of evaluating and providing feedback to teachers effectively ignores *most* of what teachers actually do to generate student growth in learning. The implication is that there would seem to be scope for using item-level data to provide more helpful feedback to teachers and students, for rethinking the ways we allocate teachers and students to classrooms, and making all manner of other important education decisions. We explore these possibilities further in the sections that follow.

6 Is the Discarded Information Meaningful?

We have shown that there is additional information in the item-level data even conditional on a single summary measure of student ability or academic performance. Is that extra information *meaningful*? That question itself raises a different issue: Meaningful for *what*? We answer this

single latent-factor null with distributions of noise and teacher value-added calibrated to match the relevant moments of our data. To be precise, we assume each teacher’s item level value-add is equal to their average value-add, and then we add noise to each item value-add where the noise distribution is assumed to be normal and with variance equal to the level of noise implied by the item-level standard errors. See appendix A.2

³⁰In addition to the simulation study and the formal test, we also probe robustness of our results on rank-swapping for teachers using the rank correlation measurement error correction from Kitagawa et al. (2018). See appendix A.8.

³¹Formally, we define information loss for teachers identically to how we define it for students, as the share of all the variation in the items that can be predicted by teachers which is ignored by focusing exclusively on the mean:
$$\text{Information loss (\%)} = \frac{\text{Predictable variation around the mean}}{\text{Total predictable variation}} = \frac{\text{var}(\delta_{jq} - \delta_j)}{\text{var}(\delta_{jq})}$$
. As with the student-level version of this exercise, in practice, we estimate the numerator and denominator necessary to construct the information loss metric using a simple variance decomposition of the raw item data. See appendix A.3 for more details.

³²We do this using a similar logic to the student level version of this test. Observe that estimating teacher value-added on the average score is equivalent to imposing a linear restriction ($\delta_{jq} = \delta_j \forall q$) on equation 8. Thus, as with the student level version of this exercise, testing a null of no information loss is equivalent to a simple test of nested linear models (Wooldridge, 2010).

by returning to first principles—the purpose of data and information is to inform *decisions*. Few people seem to think the goal of education is to literally narrowly improve test scores; after all, “teaching to the test” is far more frequently used as an epithet than as an encouragement. The information captured by tests is instead thought of as a proxy for some more general human capacity that we hope schooling develops. Is the extra information captured by the item-level data relevant for the development of those broader capacities?

In what follows, we explore a range of different outcomes that could plausibly be used as indicators of those capacities and hence serve as the sort of criteria that might guide education decisions. These include medium-term outcomes like subsequent student performance on academic coursework and disciplinary infractions, to longer-term outcomes like high school graduation, college attendance or even adult earnings. To determine whether the destroyed item-level information is meaningful, we ask whether that information helps predict these candidate decision criteria relative to relying on a single summary measure of ability.

The answer is clearly yes. We show that the extra information captured by the item-level test data is relevant for predicting decision-relevant student outcomes, and that individual test items are not all equally important for predicting these outcomes. The implication is that any decision rule that relied solely on a single aggregate ability measure would inadvertently wind up prioritizing resources or incentives (positive or negative) to many of the wrong people, and would fail to prioritize resources for many of the right people.

We also show that there is no single aggregator that is optimal for every potentially decision-relevant outcome. It turns out that different items carry different amounts of information about different candidate decision-relevant outcomes. That is, some items are relatively more important in predicting, say, course grades and disciplinary actions, while others have more predictive signal for high school graduation, while yet other items are particularly relevant for predicting college attendance. One implication is that using test data to inform education decisions will unavoidably require some normative decisions about how different outcomes should be prioritized: That is, what *are* the goals of education?

6.1 Decisions about students

Recall from section 3 that an aggregator $a^*(\cdot)$ is optimal if policy choices made using predictions of the decision criteria that is unobserved (or imperfectly observed) at the time of the decision (Y) minimize prediction errors. In other words, that:

$$MSE(Y, a^*) \leq MSE(Y, \tilde{a}) \tag{9}$$

for any alternative aggregator $\tilde{a}(\cdot)$.

We test this implication by asking whether the extra structure in the items can help us better predict who is on track for success with respect to relevant decision criteria. If we can find an alternative aggregator of the items that generates improvements in predictive power, this would imply there is scope for using the hidden structure to improve upon the existing model of decision making. To carry out this test, we estimate models of the form:

$$Y_i = F_g(D_{ig}, c_{ig}) + u_{ig} \quad (10)$$

$$Y_i = G_g(A_{ig}, c_{ig}) + v_{ig} \quad (11)$$

where Y_i is the decision criteria of interest; D_{ig} is the vector of item responses student i gave to the grade g tests, which include both ELA and math items; $A_{ig} = (a_r(D_{ig}), a_m(D_{ig}))$ are the English language arts (a_r) and math (a_m) aggregate scores derived from the relevant grade level items; c_{ig} is the student’s cohort (i.e. the calendar year student i attended grade g);³³ and F_g and G_g are grade-specific functions to be learned from the data. Our goal is to compare the predictive power of these models. Model (10) allows us to quantify the predictive power of the items.³⁴ Model (11) allows us to quantify the predictive power of the aggregates.

To learn the functions F_g and G_g , we use an off-the-shelf implementation of the gradient-boosted tree algorithm (Chen et al., 2016).³⁵ For D_{ig} , our preferred specification uses the exact answers chosen by students on each question for both the reading and math tests in a given year; however, we obtain similar results using a binary, right/wrong categorization (see appendix A.9). To account for over-fitting, we divide our data into a 75% – 25% train-test split, and calculate all of our fit metrics on the hold-out sample. Unless otherwise indicated, we estimate our predictive models for all students we can connect to the relevant candidate decision criterion of interest (see appendix table A.5), and allow the functions to be “grade-specific” by estimating the models separately for

³³Because the actual items contained in each test are different across cohorts, including the cohort variable is necessary for “item 1 on the grade 4 exam in 2015” to have a different relationship with the outcome than “item 1 on the grade 4 exam in 2016,” and, symmetrically, for “average math score on the grade 4 exam in 2015” to have a different relationship with the outcome than “average math score on the grade 4 exam in 2016.” However, to ensure that any gains in predictive power are purely measuring additional information in the items, rather than the ability of the items to better interact with the cohort variable and detect changes over time, we remove all time series variation from our fit metrics (described later on in this section) by calculating R-squared within grade-cohort (so holding c_{ig} constant) and then averaging across cohorts for a given grade-level.

³⁴Note that we can view the predicted values from this model as an alternative aggregator (as in Bond and Lang, 2013, 2018; Nielsen, 2019) and hence an out of sample comparison will directly test the relevant implication of the single latent factor model of learning.

³⁵Decision trees are widely used in machine learning to help capture high levels of interactivity among candidate predictors or “features” in predicting the outcome of interest. The problem with trees is that their structure can be quite sensitive to statistical noise in the data. The “random forest” model solves that problem by building multiple trees from different subsets of the data and averaging them together. Gradient-boosted trees are closely related but build the trees sequentially, where each tree focuses on improving predictive power in that part of the data where previous trees do not do very well. For more details see, for example, Natekin and Knoll (2013).

each grade stacked across all cohorts.³⁶

For the aggregated metrics (A_{ig}), we use simple average scores in ELA and math in the indicated grade. In many cases, this choice winds up *understating* the value of the extra information in the item-level data, in the sense that the gains are even larger when we use other common educational aggregators such as scaled scores (see appendix A.9).

We measure gains in predictive power using the difference in out-of-sample R-squared between the models trained on test items versus aggregates. We benchmark the difference by scaling the gains in predictive power by the R-squared of the model trained on aggregates. Thus, our results communicate how much additional predictive power we gain for decision making relative to a baseline of using only aggregates.

Figure 5 displays the results. In the figure, “Class failure” refers to results from models where the outcome is an indicator for failing a class in the indicated grade. “Disciplinary action” refers to results from models where the outcome is an indicator that takes a value of one if the student had any disciplinary action in the subsequent year. “High school graduation” refers to results from models where the outcome is an indicator for ever graduating high school. “College attendance” refers to results from models where the outcome is an indicator for ever attending college. “Wages” refers to results from models where the outcome is earnings 5 years after on-time graduation; however, we caution that these results come from a smaller sample since we are only able to link to earnings for a single cohort of students.³⁷

[Figure 5 about here.]

We find non-trivial gains in predictive power. On average, across grade levels, we find that the items provide a 4.7% gain in our ability to predict class failure (min = 0.0%, max = 10%); a 12% gain in our ability to predict disciplinary violations (min = 9.3%, max = 17.5%); a 19% gain in our ability to predict high school graduation (min = 13.4%, max = 24.8%); a 6.5% gain in our ability to predict college attendance (min = 5.8%, max = 7.7%) and a 44.5% gain in our ability to predict earnings. To the extent to which any of these measures are the right criteria, these results suggest there is real potential to use item-level data in a way that could substantially improve educational decisions.

To address the possibility that the results are a product of sampling variation, we formally test the null hypothesis of “no-information gain.” As suggested by equation 9, this amounts to testing the

³⁶In practice, stacking across cohorts is necessary for us to have a large enough sample to adequately tune the model hyper-parameters that govern regularization to prevent over-fitting.

³⁷While the other outcomes are measured for longer periods of time in our data, and over many cohorts, the results for earnings are based on a single cohort of 8th graders and hence leverage a substantially smaller sample. Depending on the grade, our sample sizes range from $n = 114,761$ to $320,374$ for high-school graduation, $57,978$ to $258,497$ for college attendance, and only $39,284$ for earnings. This is because the linkages necessary to measure earnings 5 years after on time high-school graduation do not go back far enough in time. This raises the possibility that the results for earnings as a candidate decision-making criterion could be due at least partly to noise in the data, rather than true signal.

null hypothesis that mean squared error on the hold-out sample is lower when using the predicted value built from the items relative to the predicted value built from the aggregates. In other words, $H_0 : MSE(Y, G) < MSE(Y, F)$ which implies a simple F-statistic (Wooldridge, 2010). We reject the null of no information gain (see Appendix Table A.7 for formal test results).

Perhaps a more intuitive way to see what the item-level information adds comes from looking at how the item-level versus aggregate models rank-order students, as in Figure 6. This not only helps us see the data in a different way, but is also useful because of how many education decisions have the flavor of rank-ordering problems (for example, remedial supports for the students at the bottom of the distribution, etc.). For the prediction of a given student outcome, we plot the predicted values from the model that uses aggregate test data to form the prediction (X-axis) versus uses the item-level test data (Y-axis). For visual clarity, we standardize the predicted values to have mean zero, standard deviation 1 prior to plotting.

Up to measurement error, the single-latent-factor model implies that for a given outcome the aggregate and item-level predictions plotted in Figure 6 should all lie *exactly* on the 45 degree line. Panel (a) shows that for the outcome of failing a math class, with the wide dispersion in the data suggesting that the single-latent-factor model does not fit the data. The other panels replicate the analysis for failing an ELA class (panel b), student disciplinary actions (panel c), high school graduation (panel d), college attendance (panel e) and earnings (panel f).

[Figure 6 (panels a-b) about here.]

[Figure 6 (panels c-d) about here.]

[Figure 6 (panels e-f) about here.]

To formally test whether the points in figure 6 lie on the 45 degree line, we regress the predicted values formed using aggregates on the predicted values formed using items and test whether the slope coefficient equals 1. To account for measurement error in the right-hand side variable, we estimate the model using a split sample two-stage least squares procedure. We can reject that the slope equals one for each outcome at every grade level (see appendix A.10). The fact that the points do not lie on the 45 degree line suggests scope for large disagreements regarding which students should be prioritized for different types of resources, interventions, remedial support, etc. depending on whether we use the benchmark, aggregate achievement measure or patterns of performance in the underlying items.

Table 4 shows the degree of disagreement between item-level and aggregate-level decision-making about which students are in the bottom 5% of the distribution.³⁸

³⁸To be precise, we calculate the disagreement share by first counting the total number of students that fall into the bottom 5% of the distribution of predictions made using either items or aggregates (but not both), and then divide by the total number of students that fall into the bottom five percent of at least one of the two distributions. We follow the approach from Petek and Pope (2023) and Rose et al. (2022) by standardizing the predictions to each have mean 0 and standard deviation 1, and then counting a student in the bottom 5% of the relevant distribution if the value falls below -1.645. This approach will be exactly correct under the assumption that the predicted values are normally

[Table 4 about here.]

Disagreement between the models formed using item-level versus aggregate-level data average 52.5% when it comes to deciding which students are at risk of not graduating high school; 48.9% for those at risk of not going to college; 52.1% for those at risk of a disciplinary infraction; 46.1% for those at risk of class failure; and 62.8% of cases for those at risk of falling into the bottom of the earnings distribution. In fact, in nearly half of all grade-year-outcome combinations, the models disagree on the *majority* of students when it comes to deciding who is in the tail of the distribution. To account for the potential of these results to be driven by sampling uncertainty, we formally test the null of no disagreement using a parametric bootstrap procedure that accounts for estimation error in the random forest model-fitting step.³⁹ We can reject the null of no disagreement for every grade-by-outcome combination (see the p-values in the 2nd to last column of Table 4). Consistent with the idea that the item-level predictions are more accurate overall, the observed high-school graduation rates are closer (i.e. more aligned with the predictions) to the item-based model in cases where the models disagree about how the student should be classified (see appendix exhibit A.6).

6.2 Decisions about Teachers

Our results for students demonstrate that there is extra information in the test items that is both relevant to educational decision making and that gets destroyed by aggregation. But what drives this extra variation is critical for understanding its broad utility. One possibility is that the additional variation is capturing certain types of skills and abilities important for student outcomes. A less generous interpretation is that the correlation with longer-run student outcomes is driven by the ability of the machine learning algorithm to use the information in the items to infer other relevant student-level characteristics. For example, imagine that high-income and low-income students with the same average score tend to answer different items correctly; the extra variation in the item-level data in that case may simply be inferring the socioeconomic status of students.

To address this concern, we turn next to teachers, where we can explore whether this pathway drives our results. Teacher value-added is constructed by residualizing test scores on observable characteristics – such as race/ethnicity, gender, and proxies for socioeconomic status – at both the student and classroom levels, ensuring that these features are mechanically uncorrelated with the measure. Moreover, if teacher value-added captures an average causal effect, as argued by [Chetty et al. \(2014a\)](#); [Koedel and Rockoff \(2015\)](#); [Bacher-Hicks et al. \(2019\)](#), it should also be orthogonal to unobserved characteristics of the student’s they teach, implying that the gains from using item-level data are not driven by the algorithm inferring the characteristics of the students in

distributed, and has the advantage that it simplifies inference considerably relative to defining the bottom 5% using the corresponding empirical quantile. However, we find similar point estimates using a threshold that is based on the empirical 5th quantiles (see appendix A.11).

³⁹See appendix A.12 for more details on the bootstrap procedure.

their classroom.⁴⁰

We estimate teacher value-added on the predicted values from the machine learning model described above, where \hat{Y}_{ig} is a predicted value from either the item model (\hat{F}) or the aggregate model (\hat{G}). Intuitively, the predicted values from the item-based machine learning model “re-weight” the items according to their importance for predicting long-run outcomes. Thus, any substantive points of disagreement about teacher value-added on the item-based model versus the model based on aggregates must be due to the fact that some teachers are systematically better at improving student performance on the items that matter most for the long-run decision of interest.

More precisely, we use the method developed in [Chetty et al. \(2014a\)](#) to estimate jack-knife, empirical-Bayes versions of the following teacher value-added specification:

$$\hat{Y}_{ig} = \delta_{j(i,g)} + \pi X_i + u_{ig}, \quad (12)$$

where $\delta_{j(i,g)}$ is a fixed effect for the teacher (j) assigned to student (i) in grade (g); and X_i is the standard vector of value-add controls (described in full detail in [appendix A.5](#)).

[Figure 7](#) shows how, relative to relying on the single aggregate test-score summary, the extra information in the test-score items leads us to rank-order teachers differently by their value-added for a given predicted outcome or decision criterion. In each sub-figure, the y-axis plots teacher-value add on predicted values formed using items, and the x-axis plots teacher value-add on predicted values formed using aggregate scores. See [section 6.1](#) for a discussion of the predictive model estimation.

[[Figure 7](#) (panels a-b) about here.]

[[Figure 7](#) (panels c-d) about here.]

[[Figure 7](#) (panel d) about here.]

As with the student results, the wide dispersion we see in the data—with so many aggregate-data and item-level TVA estimates lying off the 45 degree line—is further evidence against the single-latent-factor model. We see this for every outcome: math class failure ([panel a](#)), ELA class failure ([panel b](#)), disciplinary actions ([panel c](#)), high school graduation ([panel d](#)) and college attendance ([panel e](#)).⁴¹ To account for

⁴⁰Specifically, if we assume the teacher value-added model we develop in this section ([equation 12](#)) is causal in the sense that $\mathbb{E}(\hat{Y}_{ig} | D_{j(i,g)}, X_i) = \delta_{j(i,g)} + \pi X_i + \mathbb{E}(u_{ig} | D_{j(i,g)}, X_i) = \delta_{j(i,g)} + \pi X_i$, and where $D_{j(i,g)}$ is an indicator that takes a value of 1 if teacher j is assigned to student i in grade g and all other variables are as defined later on in this section, then it follows immediately that $cov(\delta_{j(i,g)}, X_i | D_{j(i,g)}) = 0$ and hence teacher value-added is uncorrelated with the observable characteristics of the students assigned to them. A similar argument holds for unobservable characteristics: simply decompose $u_{ig} = U_{ig} + v_{ig}$ where v_{ig} is mean zero noise and hence U_{ig} governs all other unobserved determinants of \hat{Y}_{ig} . An identical argument goes through.

⁴¹Data constraints prevent us from estimating teacher value-add on the predicted values for wages. Since we can only link one cohort of eight graders to wages 5 years after high school graduation, we cannot estimate the predicted values in other years that would be necessary to account for the classroom component of teacher-value-add variance using the standard, jack-knife empirical-Bayes method ([Chetty et al., 2014a](#)).

sampling variation, we formally test whether teacher value-add from the aggregate model predictions are comparable to teacher value-add on the item model predictions *on average* using a split sample, two-stage least squares procedure that is similar to the one deployed in section 6.1. We can reject the null hypothesis that the slope equals one at every grade level and for every outcome (see appendix A.10). As with the student results, teacher value-add formed using the individual test items and their connection to decision-relevant outcomes are systematically different than those made using just the average test score.

Table 5 gives the share of disagreement about which teachers are in the bottom 5% of the value-added distribution for different outcomes, between the value-added model using items versus aggregate scores. Because we have many fewer teacher-level observations in a given grade-year (compared to student-level observations), we present results for teachers pooled across all grades within an academic year. So, each row represents results for teacher value-add calculated on the indicated outcome. Disagreement shares for teacher value-add on predicted graduation average 44.9%; for predicted college attendance they average 39.6%; for predicted disciplinary infractions they average 42.4%, and for predicted class failure they average 51.6%. The bootstrap p-value shown in the second to last column of table 5 shows that we can definitively reject the null of no disagreement for each candidate outcome.⁴²

[Table 5 about here.]

The fact that we see similar results for this teacher value-added analysis as for student outcomes confirms that aggregating test score data is relegating decision-relevant information to the cutting-room floor—the student results do not merely reflect the fact that the item-level information is a stand-in for student socio-demographics. And the magnitudes here suggest the extra meaningful information in the item-level data has potentially important implications for hiring, firing, promotion, salary increases, extra professional development help, bonuses to work in hard-to-serve schools, etc.

6.3 Is there one optimal aggregator for every outcome?

We have shown that the extra information in the item-level data, when compared to an aggregator like the test score average or IRT ability parameter, is relevant for the student outcomes that guide education decisions. Put differently, we can predict those outcomes more accurately using the item-level data than the average or IRT parameter. Here we show that the deeper problem is not that the test average or IRT parameter is the wrong aggregator, but rather that we are aggregating at all.

Once we recognize that the individual items have information about outcomes even conditional on the average, we can see there is no reason to think that the extra information each item has need be the same for each possible outcome. In fact, we show that it is not, which in turn implies that there

⁴²See appendix A.12 for more detail on the procedure we use to generate the bootstrap p-value.

is *not* an “optimal aggregator” for every outcome or decision criterion.

We demonstrate this by showing the amount of disagreement when we use the item-level information to predict two different outcomes. Figure 8 provides an empirical illustration. We calculate predicted values from the item-level machine learning model (equation 10) with *college attendance* as the outcome, plotted against the predictions of identical models but now using *high school graduation* as the outcome.

[Figure 8 about here.]

The results shown in Figure 8 imply that different test items carry different amounts of signal about different outcomes. As discussed in section 6.1, the single latent factor model of decision making implies that the optimal aggregator should minimize mean-squared error and hence all of these points should fall *exactly* on a curve representing the mean relationship between the two predicted values.⁴³ A simple way to see this is to note that the single latent factor model implies that ranks should be preserved after optimal aggregation. But we can clearly see that is not the case — the wide degree of dispersion in the figure shows that there are clearly many students whose ranking, for the purpose of decision making, is highly dependent on the particular choice of aggregator.

Another way to see this is that the flat part of the curve at the left implies there are many students whose pattern of test-item responses suggest they have a similarly low predicted rate of college attendance, but vary enormously in their odds of high school graduation. Similarly, the steep part of the curve at right implies that there are many students who have the same (high) predicted odds of graduating from high school but, given the specific pattern of items they answered correctly, vary substantially in their likelihood of attending college. The only way for patterns like the one in Figure 8 to arise is if different questions carry different information about the long-run outcomes that we might use to guide education decisions.

We can also see this point directly in Figure 9. To build this figure, we estimate regressions of the following form:

$$Y_i = \alpha + \beta d_{iq} + \omega_{\bar{D}-q} + e_i \quad (13)$$

Where Y_i is the long-run outcome of interest (high school graduation or college attendance); d_{iq} is an indicator that takes a value of 1 if student i answered question q correctly; $\bar{D}_{-q} = (|T| - 1)^{-1} \sum_{q \in T_{-q}} d_{iq}$ is the leave- q -out average performance on the test; and $\omega_{\bar{D}-q}$ represents fixed

⁴³Observe that with infinite data the predicted values from the random forest model will be equivalent to the conditional expectation function and, since the CEF minimizes squared error, will be an optimal aggregator for each respective outcome. If there is a single latent factor, then conditioning on one optimal aggregator is equivalent to conditioning on the single latent factor, and the particular pattern of item responses should not generate systematic variation (i.e. variation around the conditional mean) in the other optimal aggregator.

effects for every possible value of the leave-out average. In practice, we estimate this model separately for each item on the 2016, 8th grade math and English language arts tests.

The parameter of interest in these models is the slope coefficient β . It tells us the predictive power of item q for high school graduation or college attendance among students with otherwise identical average performance (equivalent to ability under the single latent factor model), which is captured by the fixed effect $\omega_{\bar{D}-q}$. The x-axis of figure 9 plots the slope coefficients for high school graduation. The y-axis plots slope coefficients for college attendance. Prior to estimation, we standardize the outcomes to have mean zero and standard deviation one to meaningfully compare across outcomes with different base rates. Individual points are color-coded according to item difficulty, which we measure as the share of students who correctly answered the question. Because these coefficients are estimated with every student in Texas who took the test in this grade and year (57,594 in math and 57,595 in ELA), the standard errors on the individual slope estimates are small (ranging from 0.003 to 0.015 in standard deviation units), so we omit measures of sampling uncertainty for visual clarity.

[Figure 9 about here.]

Figure 9 shows that the different items contain different predictive power for different long-run outcomes even conditional on ability. While the relationship is positive on average (revealing that, conditional on ability, the items which predict an increase in high school graduation tend to predict college going), there exists wide dispersion around this mean. In fact, there are even items where the relationship is *oppositely* signed—items where getting them correct is associated with a lower chance of graduating high school but a higher one of attending college, and vice versa.

Clearly, the assumption that all the information in an achievement test could be summarized by a single summary statistic had the hidden implication of allowing us to avoid difficult conversations about what education is for. When economists want to, for example, compare the size of the black-white achievement gap at different ages, we anchor test scores in some long-term outcome. But our results suggest the choice of anchoring outcome is not innocuous and will implicitly up-weight or down-weight the importance of certain types of learning. Such choices are even more important for education policy. Should we hire and fire teachers based on their ability to help students get better course grades next year? Or pass to the next grade? Or graduate high school? Or attend college? All else equal, we would like to achieve all of these outcomes. But our results suggest there may be tradeoffs.

7 Is item-level information *practically* useful?

The results we have presented so far show that there is a great deal of extra information in the item-level test data that is not captured by a single aggregate measure like a test-score average or IRT ability parameter: The items capture additional signal about which students are off track for

different short-, medium- and long-term outcomes, or which teachers are more or less helpful in promoting such outcomes for their students. In this section, we discuss whether that extra information could be made *practically* useful for both education researchers and policymakers.

For informing education decisions about some current cohort of students or teachers, we first need some way to understand what each student or teacher’s pattern of item-level results implies for future outcomes that have not yet been observed for those students or teachers. That requires using the pattern of item-level results for some *past* cohort of students or teachers, for whom we can observe the outcome of interest, to estimate the item-to-outcome relationship. With those estimates in hand, we can then make some judgments about which current students or teachers to prioritize for extra help, which teachers to fire versus promote or give special merit raises, etc.

This type of practical application requires that the item-to-outcome relationships have some stability across cohorts. It also requires some way to group test items in some meaningful way, since the specific items represented on the tests taken by past cohorts of students will be different from the items on the current cohort’s test. Which items on each year’s test are “like” which items on another year’s test?

In what follows we provide an illustrative demonstration that these practical challenges can be solved and the item-level information can be useful in some feasible applications. We do *not* claim that our “proof of concept” implementation is necessarily optimal. It is very possible that future research comes up with better ways to, for example, group items. So, the results presented here should be viewed as a lower-bound for the benefits of using item-level information.

7.1 How to Group Items?

Our results show there is additional information in the item-level data that is relevant for educational decision-making, beyond what the single factor model would imply. But any feasible use of this insight will require *some* sort of grouping of items, since the items found on one year’s test are always different from the items on another year’s test.

Panels (a) and (b) in figure 10 highlight the problem. Panel (a) displays question 5 on the grade 8 test administered in year 2016, while panel (b) displays question 5 from the grade 8 test administered in 2017. Question 5 from the 2016 test involves a diagram and asks students to find the slope of a line, while question 5 from the 2017 test combines geometry and algebra in the context of finding the vertex of a triangle. It is unclear from looking at the text alone whether a student that struggles with the skills necessary to answer question 5 on the 2016 exam would also struggle with skills necessary to answer question 5 from the 2017 exam. Thus, even if we could learn the relationship between performance on item 5 and (say) high school graduation for a cohort where graduation is observed, there is no reason to believe that relation would apply to performance on item 5 for students in other test years who have not yet graduated. Without any additional structure, it is *as if* the tests had totally disjoint sets of indices, for example with the indices on the 2016 test

given by $T_{2016} = \{1, 2, 3, \dots\}$ and the indices on the 2017 test given by $T_{2017} = \{A, B, C, \dots\}$ and hence there is no natural way to link them.

[Figure 10 about here.]

The good news is that the questions administered year-over-year are not *actually* disjoint sets. For example, looking across the span of questions contained in the 2016 and 2017 grade 8 tests, there are questions which *do* seem to be clearly linked even if they do not share the same question number. For example, Panel (c) of figure 10 displays question 39 from the 2017 test, which is nearly identical in structure and content to question 5 from the 2016 test. This suggests there may be some latent structure to the question content that will align with the actual skills driving performance on individual items—the key challenge is to find a data-driven way to discover it.⁴⁴

What is needed, in other words, is some way to group or categorize items that is most relevant for whatever decision criteria society decides (course grades, disciplinary actions, high school graduation, etc.). With that in hand it would then be possible to use the mapping between items and decision criteria identified from some past cohort to inform decisions made for current students (for whom the future outcome upon which we are making the decision is not observed, but the item-level test performance is).

We offer here an “existence proof” that it is possible to develop categories that link conceptually “like” items across test years, without claiming that this is the optimal way to categorize and fully utilize the information contained in the items. Our idea is to leverage the fact that most teachers work in the same grade for multiple years, so they have classrooms of students who take the same grade-level test across different years. Intuitively, we look for the set of items that a given teacher’s students tend to do well on year after year.

To see the intuition, suppose the students assigned to a given 4th grade math teacher in the 2016-17 academic year perform well on questions 3, 4, and 5, but not on questions 1 and 2; but that the new students assigned to the teacher the next year (2017-18) do well on questions A, B, and C, but not on questions D and E. However, in a different teacher’s classroom, the converse is true — they do well on 1 and 2 the first year but not 3, 4 and 5, and the second year do well on D and E but not A, B and C. This pattern suggests placing the questions into two groups representing the distinct concepts or skills that reflect the distribution of latent abilities within the classrooms that each teacher tends to teach: $S_1 = \{3, 4, 5, A, B, C\}$ and $S_2 = \{1, 2, D, E\}$. We expect the first teacher’s students to do quite well on a test comprised of questions like those in S_1 , poorly on a test of items in S_2 , and conversely for the second teacher.

Our procedure formalizes this intuition. Let $C_{jt} = \{i | j(i, t) = j\}$ denote the set of students assigned to the classroom of teacher j and calculate the average performance of the students assigned to teacher j on item q in year t as $x_{jqt} = |C_{jt}|^{-1} \sum_{i \in C_{jt}} d_{iqt}$. Let $X_{qt} = \{x_{jqt}\}_{j \in G_t}$ denote the vec-

⁴⁴In fact, the algorithm we propose in this section *does* indeed link the items contained in panels (a) and (c) of figure 10 across test years using a purely data driven approach.

tor of average classroom-level performance on item q in year t for all teachers (j) who taught a classroom in the relevant grade in that exam year (G_t). Let $S = \{S_1, S_2, \dots, S_K\}$ denote a partition of the question number by exam-year indices (i.e. qt) into K disjoint sets. To link the questions across test years, we implement a K-means clustering algorithm (Hastie et al., 2009):

$$\arg \min_S \sum_{k=1}^K \sum_{X \in S_k} \|X - \mu_k\|^2 \quad (14)$$

and where μ_k is the centroid of S_k . Observe that equation (14) finds a partition of questions by minimizing the distance between the average performance within cluster and the classroom level performance of teachers on the items contained in that cluster. Importantly, because classroom level performance for a given teacher can span multiple test-years, it is possible for the algorithm to create partitions of the item indices such that items from different test years are included in the same set.

We turn next to using this categorization procedure to carry out a feasible implementation of the item-level decision approach. To the extent to which our item categorization is sub-optimal, the result would be to *understate* the potential gains from using item-level information rather than aggregate test data to inform key education decisions.

7.2 How well does the feasible item-level implementation perform?

Figure 11 provides a schematic that illustrates our feasible implementation of the item-level approach. With our item categorization system in place to link test items across years, we take the earliest cohort of 8th graders for whom we can measure high school graduation in our dataset; this is for those students who were 8th graders in academic year 2012-13 and so, had they stayed on track in school, would have graduated high school in 2016-17. We then predict high school graduation in 2016-17 for that cohort using 2012-13 8th grade test scores in two ways, first using the test score categories discovered by our algorithm (capitalizing on the item-level information) and then using the status quo approach of using an aggregate test score measure.

[Figure 11 about here.]

We can then do a policy simulation of replacing the bottom 5% of teachers in the estimated value-added distribution, horse-racing a version using aggregate test data versus our feasible implementation of the item-level approach. The school administrators in fall 2017 know nothing about the longer-term future graduation outcomes of that year's 8th graders at the time they are making decisions about which 8th grade teachers in that year are in the bottom 5% of the graduation value-added distribution. The academic 2017-18 administrators must choose between two backward-looking options: an estimate of teacher value-added from the year 2012-13 cohort based

on the aggregate test data; and an estimate of teacher value-added from the 2012-13 cohort on high school graduation, based on the test-item-cluster from our algorithm that is most highly correlated with high-school graduation. For the item-level prediction to dominate requires that our proof-of-concept categorization for the item-level data captures some meaningful structure, and that extra structure is stable over time from the 2012-13 to 2017-18 8th grade cohorts.

We evaluate the gains of the item-level approach by capitalizing on the fact that we (the researchers) can observe actual high school graduation outcomes for the year 2017-18 8th grade cohort, even if the fall 2017 decision-makers in our exercise cannot. (We emphasize again we only use that later information to evaluate, not implement, the policy). Specifically, we:

- Learn the item-cluster-to-graduation relationship using the 2012-13 eighth grade cohort;
- Rank-order teachers by their value-add on the aggregate test score and the relevant item-cluster in the 2016-17 academic year (as if we did not see the graduation outcomes for the 2016-17 cohort);
- Replace the bottom 5% of teachers using either the feasible item-cluster approach or aggregate test scores;
- Score the two policies against one another using the fact that in our data we can actually see the high school graduation outcomes for the 2016-17 cohort in academic year 2020-21.

To evaluate the policy gain, we calculate value-add using model 12 on actual, observed high school graduation and compute the implied increase in high school graduates that would accrue if we replace the teacher who is let go with a teacher who has average value-add on high school graduation. In practice, the linkages in our data allow us to perform this policy gain exercise for 2 cohorts of 8th graders and one cohort of 7th graders.

We reiterate that for our policy simulation, the identification of the bottom 5% of teachers in all three of these cohorts is made without using any of that 2020-21 graduation information (since actual school administrators in year 2016-2017 would not know that either). In other words, the decisions we simulate are exactly constrained by the information limitations that education policy-makers would face; it is only our own *evaluation* of this policy that is assumed to have the benefit of hindsight. Figure 12 contains the results.

[Figure 12 about here.]

We find that implementing a feasible version of the item-based policy generates a 20.6% larger gain in high school graduation than the version implemented with traditional value add. In levels, we find that the item-based policy creates just under 500 additional high school graduates per grade-year statewide in Texas — nearly 100 more than the number of additional high school graduates created using traditional value-add.

To formally quantify the cost-benefit ratio, we recast it in terms of the marginal value of public funds (MVPF) (Hendren and Sprung-Keyser, 2020). We calculate the MVPF using the formula

$MPVF = \frac{(1-\tau) \sum_i \Delta Earnings_i}{Cost - \tau \sum_i \Delta Earnings_i}$, where τ is the marginal tax rate, $Cost$ is the implementation cost, and $\sum_i \Delta Earnings_i$ represents the change in present-value monetary earnings from being a high school graduate relative to a high school dropout. We assume the average employed high school dropout earns \$38,500, two-thirds of high school dropouts are employed, a 10% return to high school graduation (Card, 1999), and the federal marginal tax rate of 12%.⁴⁵ We find that the MVPF of switching to an item-based policy is infinite at any implementation cost less than \$4 million per year, and will generate a value greater than one at any cost less than \$33 million per year.

Importantly, implementing the policy at the item level, rather than using aggregate scores, would in practice entail only modest additional cost. Because the underlying data are already collected, the only incremental expenses—relative to a policy based on aggregate scores—would stem from data analysts’ time and the computing resources needed to run our algorithm. To the extent to which there are costs they would be up-front investments (e.g., developing the necessary code) that could be amortized over many years.

In our view, the annualized costs almost surely fall well under \$4 million – implying that in the classic bottom 5% thought experiment, switching to an item-based version of the policy almost surely yields a MVPF of infinity—or, put differently, exceeds the MVPF of even well-regarded interventions like Head Start, universal pre-K, and school finance equalization. The only way the use of the item-level data would yield a MVPF below 1 would be if the cost relative to using aggregate data exceeded \$30 million a year, which would represent about a third of what Texas spent annually on developing and administering the state’s standardized tests during our study period.⁴⁶

More broadly, the cost of switching almost *any* policy from an aggregate-score based decision making process to an item-based decision process will be similarly low, implying large gains, in an MVPF sense, from switching to items for any policy decision that has a non-trivial benefit.

8 Conclusion

Our results suggest the value of rethinking how we use testing data in both education research and education practice. Upstream of researchers and practitioners, before they ever get access to any testing results, there is a step that is so widely ignored as to be almost hiding in plain sight: Aggregation. Much of the information in the item-level test data winds up being thrown away and collapsed into a single scalar, usually an average or some sort of “ability” parameter from a statistical model like IRT.

⁴⁵The high school dropout earnings and employment figures come from National Center for Education Statistics (2023), adjusted to 2025 dollars. A discount rate of 3% is used and we assume the state tests six grades (grades 3-8) each year.

⁴⁶See <https://www.nbcdfw.com/news/local/90-million-tab-for-staar-testing-includes-pricey-meetings-travel-consultants/2073351/>.

We have argued here that this aggregation step cannot be assumed to be innocuous. Using item-level information from the Texas public schools, we have shown that conditional on the test average, item-level data adds information that is meaningfully and practically important for research and policy decisions about which students and teachers are doing well versus poorly, in which specific areas, and for what decision-relevant outcomes. Aggregation leads us to throw away most of the predictable variation across teachers in value-added, misclassify nearly half the students who fall in the bottom of the distribution in terms of predicted educational trajectories, and understate (and attenuate) the effects of policies like replacing the bottom 5% of teachers in the value-add distribution. Since different items wind up having different amounts of signal for different decision-relevant outcomes, aggregation has prevented us from realizing the need for a conversation about how education should prioritize the different candidate goals; what are schools for, exactly?

The key shift in perspective is to recognize that many, perhaps most, education decisions are prediction policy problems (Kleinberg et al., 2015; Mullainathan, 2025). Just like judge pre-trial decisions hinge on a prediction of a defendant’s risk (Kleinberg et al., 2018; Rambachan, 2024), decisions about which students to provide extra help or which teachers to hire and fire hinge on a prediction of future educational outcomes. Framing the problem this way makes clear we need to extract as much predictive signal from the available candidate predictors or “features” as possible, rather than compress much (or even most) of the information in the features before feeding them to the prediction model’s training algorithm.

What a future that took item-level data more seriously would look like, exactly, remains unclear and is beyond the scope of this paper to determine, since this raises some new complications beyond what we see in many other prediction policy problems. For example, some sort of categorization of items will be required to make this approach practically useful, since the exact composition of test items changes from year to year in every test. We have provided a proof-of-concept demonstration here that this is possible, although we do not claim that our approach is optimal. If our approach is sub-optimal, we understate the potential gains from making greater use of item-level data.

More generally, education is hardly alone in having upstream data aggregation that is largely ignored. The same issue arises in a wide range of other economically important areas including health, housing, transportation, crime, public finance, consumer finance and even the measurement of inflation and other fundamental economic phenomena. Evidence that aggregation of education data is not innocuous suggests there is no reason to assume aggregation is harmless in any other area of economics.

References

- Abdulkadiroğlu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, and Parag A. Pathak**, “Accountability and flexibility in public schools: Evidence from Boston’s charters and pilots,” *Quarterly Journal of Economics*, 2011, 126 (2), 699–748.
- Adukia, Anjali, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz**, “What We Teach About Race and Gender: Representation in Images and Text of Children’s Books,” *Quarterly Journal of Economics*, 2023, 138 (4), 2225–2285.
- **and Emileigh Harrison**, “Separation of Church and State Curricula? Examining Public and Religious Private School Textbooks,” Technical Report, National Bureau of Economic Research 2025.
- Agostinelli, Francesco and Matthew Wiswall**, “Estimating the technology of children’s skill formation,” *Journal of Political Economy*, 2025, 133 (3), 846–887.
- Ahmed, Ishita, Masha Bertling, Lijin Zhang, Andrew D. Ho, Prashant Loyalka, Hao Xue, Scott Rozelle, and Benjamin W. Domingue**, “Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials,” *Journal of Research on Educational Effectiveness*, 2024.
- Angrist, Joshua D. and Victor Lavy**, “Using Maimonides’ rule to estimate the effect of class size on scholastic achievement,” *Quarterly Journal of Economics*, 1999, 114 (2), 533–575.
- **, Parag A. Pathak, and Christopher R. Walters**, “Explaining charter school effectiveness,” *American Economic Journal: Applied Economics*, 2013, 5 (4), 1–27.
- **, Peter D. Hull, Parag A. Pathak, and Christopher R. Walters**, “Leveraging lotteries for school value-added: Testing and estimation,” *Quarterly Journal of Economics*, 2017, 132 (2), 871–919.
- Arcidiacono, Peter and Michael Lovenheim**, “Affirmative action and the quality–fit trade-off,” *Journal of Economic Literature*, 2016, 54 (1), 3–51.
- Bacher-Hicks, Andrew, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger**, “An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys,” *Economics of Education Review*, 2019, 73, 101919.
- Barlevy, Gadi and Derek Neal**, “Pay for percentile,” *American Economic Review*, 2012, 102 (5), 1805–1831.
- Battaglini, Marco, Luigi Guiso, Chiara Lacava, Douglas L. Miller, and Eleonora Patacchini**, “Refining public policies with machine learning: The case of tax auditing,” *Journal of Econometrics*, 2025, 249, 105847.
- Bergman, Peter, Elizabeth Kopko, and Julio E. Rodriguez**, “A Seven-College Experiment Using Algorithms to Track Students: Impacts and Implications for Equity and Fairness,” Working Paper 28948, National Bureau of Economic Research 2021.
- Bettinger, Eric P., Brent J. Evans, and Devin G. Pope**, “Improving college performance and retention the easy way: Unpacking the ACT exam,” *American Economic Journal: Economic Policy*, 2013, 5 (2), 26–52.
- Bhatt, Monica P., Jonathan Guryan, Salman A. Khan, Michael LaForest-Tucker, and Bhavya**

- Mishra**, “Can Technology Facilitate Scale? Evidence from a Randomized Evaluation of High Dosage Tutoring,” Working Paper 32510, National Bureau of Economic Research 2024.
- Biasi, Barbara and Song Ma**, “The Education-Innovation Gap,” Working Paper 29853, National Bureau of Economic Research 2022.
- Billings, Stephen B., David J. Deming, and Jonah Rockoff**, “School segregation, educational attainment, and crime: Evidence from the end of busing in Charlotte-Mecklenburg,” *Quarterly Journal of Economics*, 2014, 129 (1), 435–476.
- Blattner, Laura and Scott Nelson**, “How costly is noise? Data and disparities in consumer credit,” *arXiv preprint arXiv:2105.07554*, 2021.
- Bleiberg, Joshua, Eric Brunner, Erica Harbatkin, Matthew A. Kraft, and Matthew G. Springer**, “Taking teacher evaluation to scale: The effect of state reforms on achievement and attainment,” *Journal of Political Economy Microeconomics*, 2025, 3 (3), 568–610.
- Bond, Timothy N. and Kevin Lang**, “The evolution of the Black-White test score gap in Grades K–3: The fragility of results,” *Review of Economics and Statistics*, 2013, 95 (5), 1468–1479.
- and —, “The black–white education scaled test-score gap in grades K-7,” *Journal of Human Resources*, 2018, 53 (4), 891–917.
- Brown, Christina, Supreet Kaur, Geeta Kingdon, and Heather Schofield**, “Cognitive endurance as human capital,” *Quarterly Journal of Economics*, 2025, 140 (2), 943–1002.
- Card, David**, “The causal effect of education on earnings,” *Handbook of Labor Economics*, 1999, 3, 1801–1863.
- and **A. Abigail Payne**, “School finance reform, the distribution of school spending, and the distribution of student test scores,” *Journal of Public Economics*, 2002, 83 (1), 49–82.
- Cascio, Elizabeth U. and Douglas O. Staiger**, “Knowledge, Tests, and Fadeout in Educational Interventions,” Working Paper 18038, National Bureau of Economic Research 2012.
- Chen, Tianqi, Bing Xu, Chiyuan Zhang, and Carlos Guestrin**, “Training deep nets with sub-linear memory cost,” *arXiv preprint arXiv:1604.06174*, 2016.
- Chetty, Raj, John N. Friedman, and Jonah E. Rockoff**, “The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood,” Working Paper 17699, National Bureau of Economic Research 2011.
- , —, and —, “Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates,” *American Economic Review*, 2014, 104 (9), 2593–2632.
- , —, and —, “Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood,” *American Economic Review*, 2014, 104 (9), 2633–2679.
- Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan**, “A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions,” in Sorelle A. Friedler and Christo Wilson, eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Vol. 81 of *Proceedings of Machine Learning Research* PMLR 2018, pp. 134–148.
- Coleman, James S.**, “Equality of Educational Opportunity: Summary Report,” 1966.
- Cullen, Julie Berry, Brian A. Jacob, and Steven Levitt**, “The effect of school choice on partici-

- pants: Evidence from randomized lotteries,” *Econometrica*, 2006, 74 (5), 1191–1230.
- Cunha, Flavio and James J. Heckman**, “Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation,” *Journal of Human Resources*, 2008, 43 (4), 738–782.
- , – , and **Susanne M. Schennach**, “Estimating the technology of cognitive and noncognitive skill formation,” *Econometrica*, 2010, 78 (3), 883–931.
- de Ayala, Rafael Jaime**, *The Theory and Practice of Item Response Theory*, New York: The Guilford Press, 2009.
- Dee, Thomas S. and Brian Jacob**, “The impact of No Child Left Behind on student achievement,” *Journal of Policy Analysis and management*, 2011, 30 (3), 418–446.
- Deming, David J.**, “Using school choice lotteries to test measures of school effectiveness,” *American Economic Review*, 2014, 104 (5), 406–411.
- DePaepe, James, Daniel Matthews, Kirk Mathias, Stephanie Harris, Sigrid Davison, Elizabeth Lee, and David Braskamp**, “CWU teacher time study: How Washington public school teachers spend their work days,” Technical Report, Central Washington University 2015.
- Ding, Weili, Yipeng Tang, and Yongmei Hu**, “Closing the gender gap in science: new evidence from urban China,” *Education Economics*, 2023, 31 (5), 531–554.
- Dynarski, Susan, Daniel Hubbard, Brian Jacob, and Silvia Robles**, “Estimating the Effects of a Large For-Profit Charter School Operator,” Working Paper 24428, National Bureau of Economic Research 2018.
- Figlio, David and Susanna Loeb**, “School accountability,” *Handbook of the Economics of Education*, 2011, 3, 383–421.
- Gilbert, J. B., Z. Himmelsbach, J. Soland, M. Joshi, and B. W. Domingue**, “Estimating heterogeneous treatment effects with item-level outcome data: Insights from Item Response Theory,” *Journal of Policy Analysis and Management*, 2025. Forthcoming.
- Gilbert, Joshua B., Luke W. Miratrix, Mridul Joshi, and Benjamin W. Domingue**, “Disentangling Person-Dependent and Item-Dependent Causal Effects: Applications of Item Response Theory to the Estimation of Treatment Effect Heterogeneity,” *Journal of Educational and Behavioral Statistics*, 2024, 50 (1), 72–101.
- Grimon, Marie-Pascale and Christopher Mills**, “Better together? A field experiment on human-algorithm interaction in child protection,” *arXiv preprint arXiv:2502.08501*, 2025.
- Guryan, Jonathan, Jens Ludwig, Monica P. Bhatt, Philip J. Cook, Jonathan M. V. Davis, Kenneth Dodge, George Farkas, Jr. Fryer Roland G., Susan Mayer, Harold Pollack, Laurence Steinberg, and Greg Stoddard**, “Not Too Late: Improving Academic Outcomes among Adolescents,” *American Economic Review*, 2023, 113 (3), 738–65.
- Hambleton, Ronald K., Hariharan Swaminathan, and H. Jane Rogers**, *Fundamentals of item response theory*, Newbury Park, CA: Sage Publications, 1991.
- Hanushek, Eric A.**, “The economics of schooling: Production and efficiency in public schools,” *Journal of Economic Literature*, 1986, 24 (3), 1141–1177.
- , “Some findings from an independent investigation of the Tennessee STAR experiment and from

- other investigations of class size effects,” *Educational Evaluation and Policy Analysis*, 1999, 21 (2), 143–163.
- , “Teacher Deselection,” in Dan Goldhaber and Jane Hannaway, eds., *Creating a New Teaching Profession*, Washington, DC: Urban Institute Press, 2009, pp. 165–180.
 - **and Steven G. Rivkin**, “Teacher quality,” *Handbook of the Economics of Education*, 2006, 2, 1051–1078.
 - , **Jin Luo, Andrew J. Morgan, Minh Nguyen, Ben Ost, Steven G. Rivkin, and Ayman Sha-keel**, “The Effects of Comprehensive Educator Evaluation and Pay Reform on Achievement,” Working Paper 31073, National Bureau of Economic Research 2023.
- Hashim, Shirin A., Thomas J. Kane, Thomas Kelley-Kemple, Mary E. Laski, and Douglas O. Staiger**, “Have income-based achievement gaps widened or narrowed?,” Technical Report, National Bureau of Economic Research 2020.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman**, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* Springer Series in Statistics, 2nd ed., Springer, 2009.
- Heckman, James J. and Alan B. Krueger**, *Inequality in America: What Role for Human Capital Policies?*, Cambridge, MA: MIT Press, 2005.
- , **Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz**, “The rate of return to the HighScope Perry Preschool Program,” *Journal of Public Economics*, 2010, 94 (1–2), 114–128.
- Hendren, Nathaniel and Ben Sprung-Keyser**, “A unified welfare analysis of government policies,” *Quarterly Journal of Economics*, 2020, 135 (3), 1209–1318.
- Hoxby, Caroline M.**, “Does competition among public schools benefit students and taxpayers?,” *American Economic Review*, 2000, 90 (5), 1209–1238.
- , “The effects of class size on student achievement: New evidence from population variation,” *Quarterly Journal of Economics*, 2000, 115 (4), 1239–1285.
 - **and Christopher Avery**, “The Missing ”One-Offs”: The Hidden Supply of High-Achieving, Low-Income Students,” Working Paper 18586, National Bureau of Economic Research 2012.
- Humphries, John Eric, Christopher Neilson, Xiaoyang Ye, and Seth D. Zimmerman**, “Parents’ Earnings and the Returns to Universal Pre-Kindergarten,” Working Paper 33038, National Bureau of Economic Research 2024.
- Jackson, C. Kirabo**, “What do test scores miss? The importance of teacher effects on non–test score outcomes,” *Journal of Political Economy*, 2018, 126 (5), 2072–2107.
- , “Does School Spending Matter? The New Literature on an Old Question,” in Laura Tach, Rachel Dunifon, and Daniel L. Miller, eds., *Confronting Inequality: How Policies and Practices Shape Children’s Opportunities*, American Psychological Association, 2020, pp. 165–186.
 - , **Rucker C. Johnson, and Claudia Persico**, “The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms,” *Quarterly Journal of Economics*, 2015, 131 (1), 157–218.
 - , **Shanette C. Porter, John Q. Easton, Alyssa Blanchard, and Sebastián Kiguel**, “School effects on socioemotional development, school-based arrests, and educational attainment,” *Amer-*

- ican Economic Review: Insights*, 2020, 2 (4), 491–508.
- Jacob, Brian A.**, “The challenges of staffing urban schools with effective teachers,” *The future of children*, 2007, pp. 129–153.
- **and Lars Lefgren**, “Remedial education and student achievement: A regression-discontinuity analysis,” *Review of Economics and Statistics*, 2004, 86 (1), 226–244.
- Jencks, Christopher and Meredith Phillips, eds**, *The Black-White Test Score Gap*, Washington, D.C.: Brookings Institution Press, 1998.
- Jr, Roland G. Fryer and Steven D. Levitt**, “Understanding the black-white test score gap in the first two years of school,” *Review of Economics and Statistics*, 2004, 86 (2), 447–464.
- Kane, Thomas J. and Douglas O. Staiger**, “The promise and pitfalls of using imprecise school accountability measures,” *Journal of Economic Perspectives*, 2002, 16 (4), 91–114.
- **and —**, “Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation,” Working Paper 14607, National Bureau of Economic Research 2008.
- Kitagawa, Toru, Martin Nybom, and Jan Stuhler**, “Measurement Error and Rank Correlations,” CeMMAP Working Paper CWP28/18, Centre for Microdata Methods and Practice (CeMMAP) 2018.
- Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan**, “Human decisions and machine predictions,” *Quarterly Journal of Economics*, 2018, 133 (1), 237–293.
- **, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer**, “Prediction policy problems,” *American Economic Review*, 2015, 105 (5), 491–495.
- Kline, Patrick and Christopher R. Walters**, “Evaluating public programs with close substitutes: The case of Head Start,” *Quarterly Journal of Economics*, 2016, 131 (4), 1795–1848.
- Koedel, Cory and Jonah E Rockoff**, “Value-added modeling: A review,” *Economics of Education Review*, 2015, 47, 180–195.
- Krueger, Alan B.**, “Experimental estimates of education production functions,” *Quarterly Journal of Economics*, 1999, 114 (2), 497–532.
- Lee, Steven and Matthew Schaeffer**, “Content Reliability and Test Score Disparities: Evidence from Texas,” 2024.
- Li, Danielle, Lindsey Raymond, and Peter Bergman**, “Hiring as exploration,” *Review of Economic Studies*, 2025, p. rdaf040.
- Loveless, Tom**, “Strengthening the Curriculum,” in C. E. Finn and R. Sousa, eds., *What Lies Ahead for America’s Children and Their Schools*, Stanford, CA: Hoover Institution Press, 2014, pp. 137–148.
- Mountjoy, Jack**, “Community Colleges and Upward Mobility,” *American Economic Review*, 2022, 112 (8), 2580–2630.
- Mullainathan, Sendhil**, “Economics in the Age of Algorithms,” *AEA Papers and Proceedings*, 2025, 115, 1–23.
- Natekin, Alexey and Alois Knoll**, “Gradient boosting machines, a tutorial,” *Frontiers in neuro-robotics*, 2013, 7, 21.

- National Center for Education Statistics**, “Digest of Education Statistics 2023,” Technical Report, U.S. Department of Education 2023.
- National Commission on Excellence in Education**, “A Nation at Risk: The Imperative for Educational Reform,” 1983.
- Nickow, Andre, Philip Oreopoulos, and Vincent Quan**, “The promise of tutoring for PreK–12 learning: A systematic review and meta-analysis of the experimental evidence,” *American Educational Research Journal*, 2024, 61 (1), 74–107.
- Nielsen, Eric**, “Test questions, economic outcomes, and inequality,” 2019.
- Petek, Nathan and Nolan G. Pope**, “The multidimensional impact of teachers on students,” *Journal of Political Economy*, 2023, 131 (4), 1057–1107.
- Rambachan, Ashesh**, “Identifying prediction mistakes in observational data,” *Quarterly Journal of Economics*, 2024, 139 (3), 1665–1711.
- , **Amanda Coston, and Edward Kennedy**, “Robust design and evaluation of predictive algorithms under unobserved confounding,” *arXiv preprint arXiv:2212.09844*, 2022.
- Reardon, Sean F.**, “The Widening Academic Achievement Gap Between the Rich and the Poor,” in David B. Grusky and Jasmine Hill, eds., *Inequality in the 21st Century: A Reader*, Routledge, 2018, pp. 177–189.
- Reckase, Mark D.**, “The difficulty of test items that measure more than one ability,” *Applied Psychological Measurement*, 1985, 9 (4), 401–412.
- Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger**, “Can you recognize an effective teacher when you recruit one?,” *Education Finance and Policy*, 2011, 6 (1), 43–74.
- Rose, Evan K., Jonathan T. Schellenberg, and Yotam Shem-Tov**, “The Effects of Teacher Quality on Adult Criminal Justice Contact,” Working Paper 30274, National Bureau of Economic Research 2022.
- Rothstein, Jesse**, “Teacher quality in educational production: Tracking, decay, and student achievement,” *Quarterly Journal of Economics*, 2010, 125 (1), 175–214.
- , “Measuring the impacts of teachers: Comment,” *American Economic Review*, 2017, 107 (6), 1656–1684.
- van der Linden, Wim J. and Ronald K. Hambleton, eds**, *Handbook of Modern Item Response Theory*, 1st ed., Springer, 1997.
- , **ed.**, *Handbook of Item Response Theory: Three-Volume Set*, CRC Press, 2016.
- Walters, Christopher R.**, “The demand for effective charter schools,” *Journal of Political Economy*, 2018, 126 (6), 2179–2223.
- Woessmann, Ludger**, “The importance of school systems: Evidence from international differences in student achievement,” *Journal of Economic Perspectives*, 2016, 30 (3), 3–32.
- Wooldridge, Jeffrey M.**, *Econometric analysis of cross section and panel data*, 2nd ed., Cambridge, MA: MIT Press, 2010.



Performance: 3rd Grade

Miguel Castro

Enrolled Grade: 3

Date of Birth: 01/01/01 Student ID: *****9367 Local Student ID: --- District: 257-999 ZY CRUSE ISD

Your Child's Performance at a Glance



Reading



Did Not Meet

Grade Level

Test Date: Spring 2018



Mathematics



Did Not Meet

Grade Level

Test Date: Spring 2018

Reading Test Date: Spring 2018



CATEGORY	ANSWERED CORRECTLY
1. Understanding Across Genres	0 of 5
2. Understanding/Analysis of Literary Texts	14 of 15
3. Understanding/Analysis of Informational Texts	0 of 14
TOTAL	14 of 34

15th PERCENTILE

Your child scored the same or better than 15% of all Grade 3 students in Texas.

Mathematics Test Date: Spring 2018



CATEGORY	ANSWERED CORRECTLY
1. Numerical Representations and Relationships	0 of 8
2. Computations and Algebraic Relationships	0 of 13
3. Geometry and Measurement	5 of 7
4. Data Analysis and Personal Financial Literacy	0 of 4
TOTAL	5 of 32

Current Quantile Measure: 598Q

1st PERCENTILE

Your child scored the same or better than 1% of all Grade 3 students in Texas.

Figure 1: Example of standardized test reports provided to parents

Notes: Figure displays an example STAAR standardized test report taken from the Texas Education Agency website (<https://tea.texas.gov/student-assessment/testing/csr/taar-report-card-example-grade-3-2018.pdf>). Reports like this, which emphasize aggregate achievement metrics, are provided to parents in Texas every year to communicate their child's academic progress.

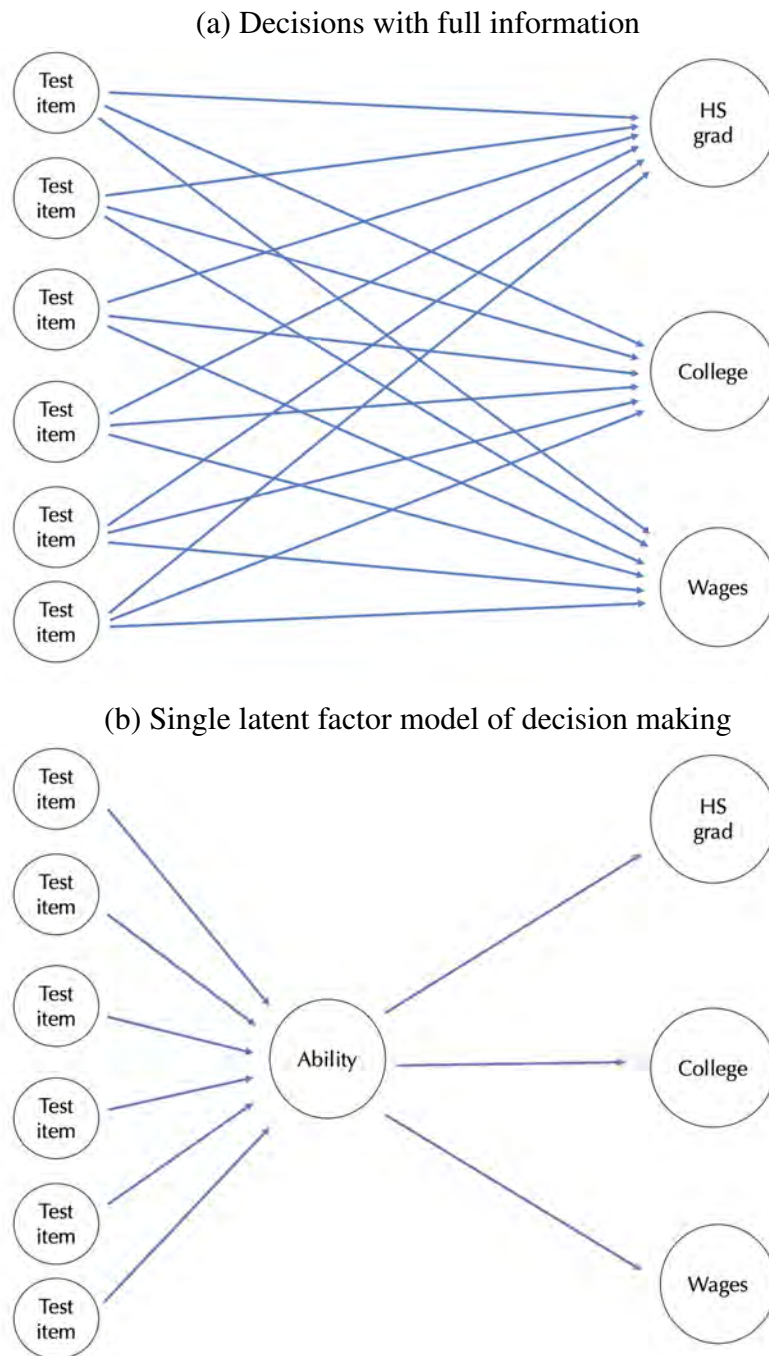
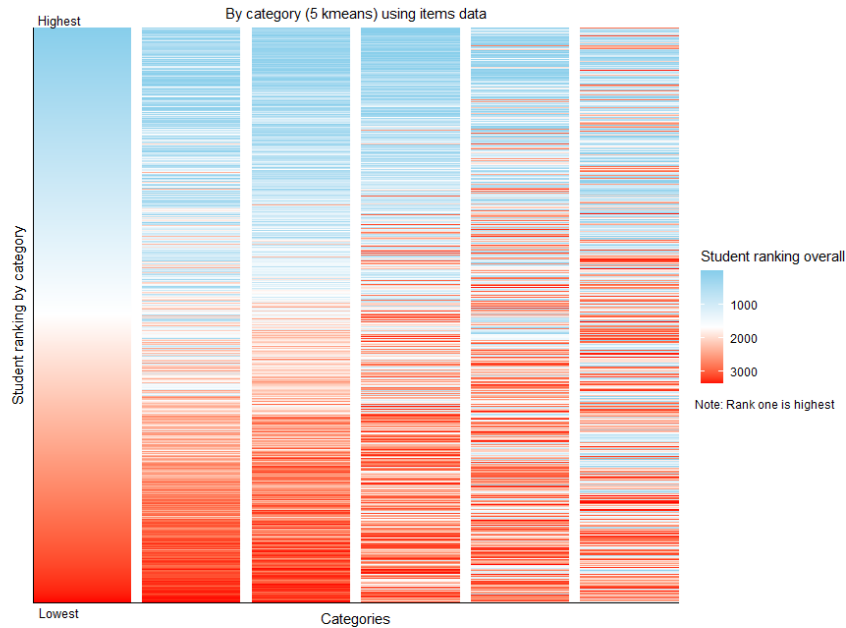


Figure 2: Two different views of decision-making with standardized tests

Notes: This figure provides an illustration of information loss when decisions are based on a single aggregate achievement measure. Panel (a) illustrates the case in which items contain independent signals for different decision-relevant outcomes, even conditional on an overall aggregate score. Panel (b) represents the common assumption underlying current practice: that all information in the test can be captured by a single underlying ability. For a formal mathematical argument, see Section 3.

(a) Distribution of student-by-category interactions



(b) Placebo distribution

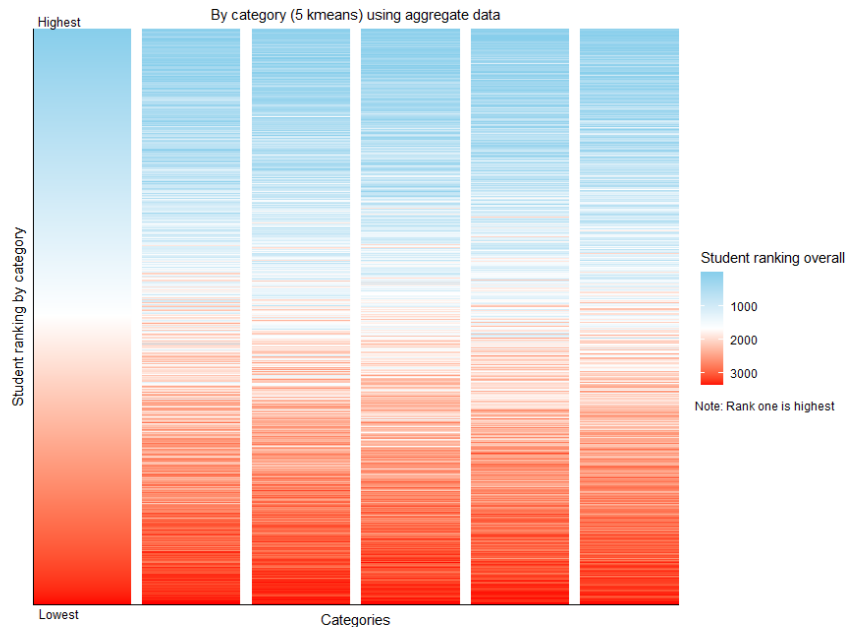


Figure 3: Hidden structure in student performance: Students who perform well on the overall test score do not perform well on every item category

Notes: Figure 3(a) takes a random sample of 1% of students taking the fourth grade test in 2016 and compares each student's rank both in terms of average test score (left-most column) and for five categories. We construct categories using k-means clustering with 5 clusters (see Appendix A.1 for alternative categorizations). The left-most column indicates a student's average performance across all questions, with each horizontal line denoting a student. The remaining five columns then re-ranks each student based on their performance in that category, but holding their color fixed based on their ranking in the left-most column. The extent of 'color mixing' therefore represents the degree to which students' ranks change from one category to another. Figure 3(b) shows how this distribution would look under solely sampling variation. We implement this placebo test by assuming that each student's performance for a given category is equal to their average score plus noise which is assumed to be normally distributed with variance set to match the empirical sampling variability (i.e. the standard error) of the student's actual category level performance.

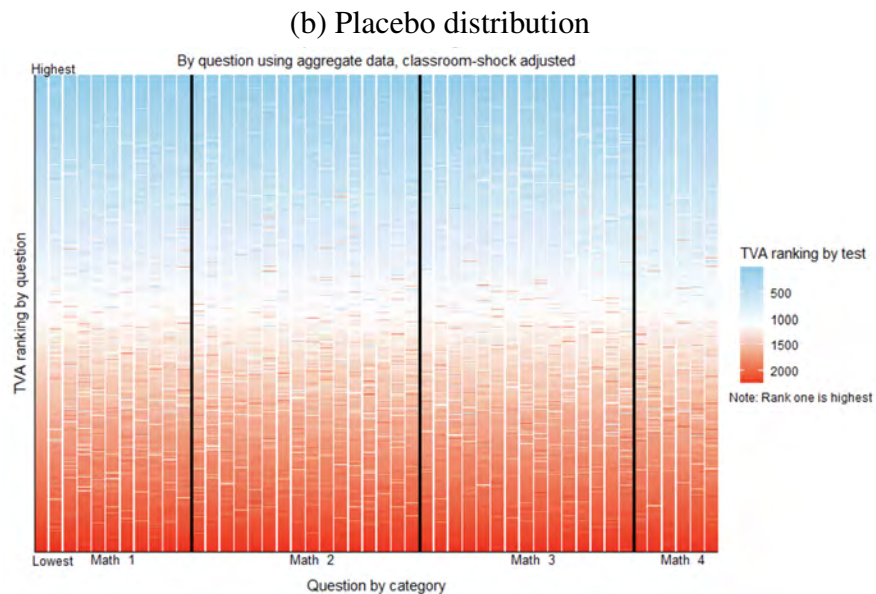
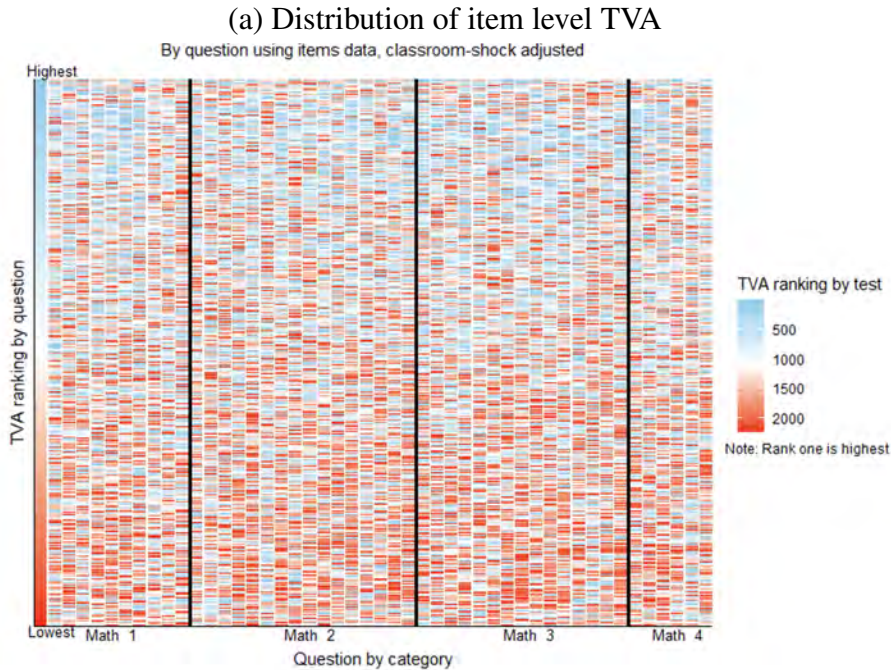


Figure 4: Hidden structure in teacher value-added: teachers with high value-add on the overall test score do not have high value-add on every item

Notes: Panel (a) of this figure takes all 2,280 fourth grade teachers in 2016 that meet our estimation criteria and compares each teacher’s rank both in terms of traditional value-add and for item-by-item value-add. In the left-most column, which corresponds to a teacher’s traditional value-add, each horizontal line denotes a teacher with their position and color determined by their rank in the value-add distribution. The remaining columns then re-rank each teacher based on their value-add on that specific item, but holding their color fixed based on their ranking in the left-most column. Thus, the extent of ‘color mixing’ represents the degree to which teachers’ ranks change from one item to another. Dark black vertical lines group individual items into state content categories. Panel (b) of this figure shows how this distribution would look if the only reason for rank-swapping were sampling variation. We implement this placebo test by assuming that each teacher’s item value-add is equal to their traditional value-add plus noise which is assumed to be normally distributed with variance set to match the empirical sampling variability (i.e. the standard error) of the teacher’s actual item-specific performance.

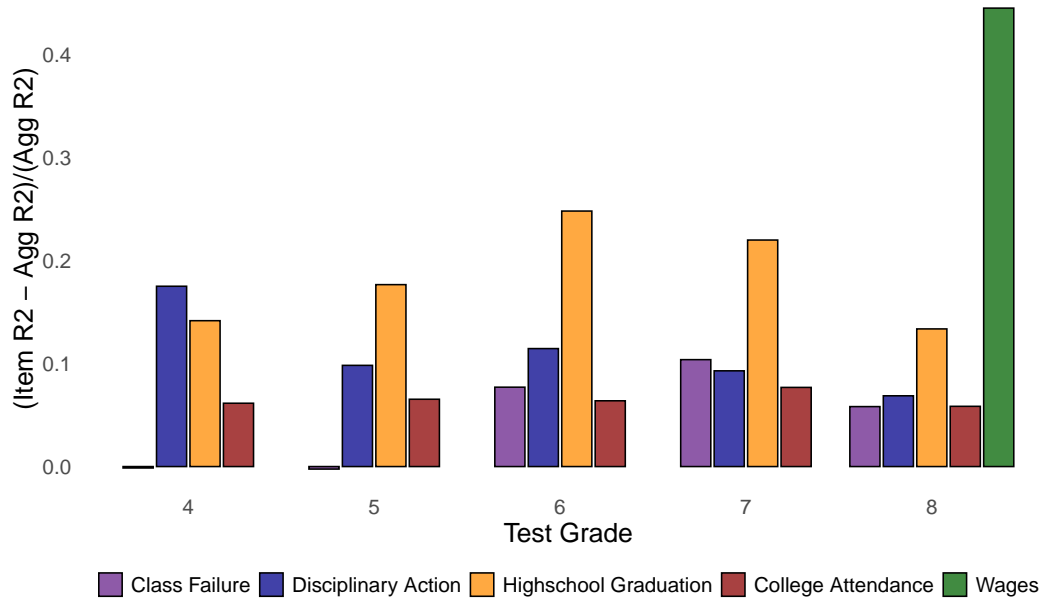
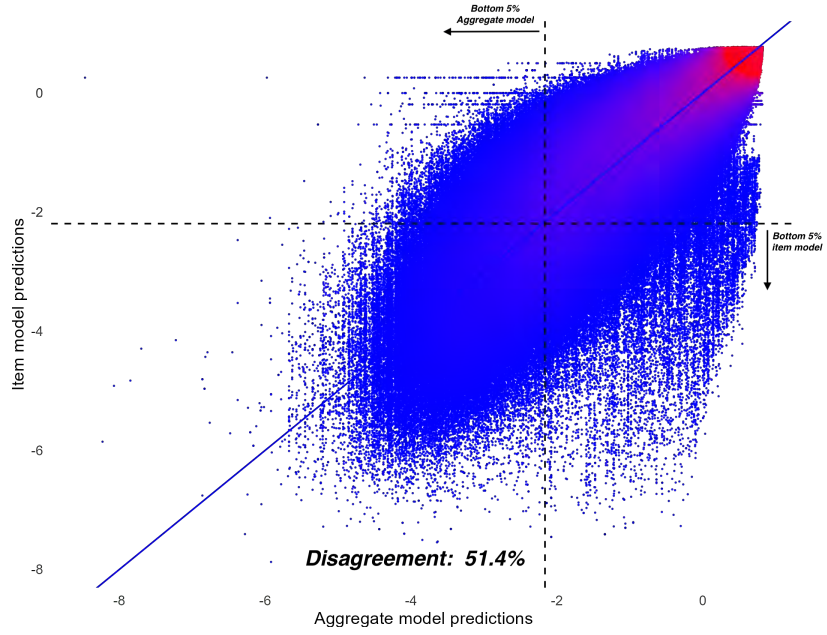


Figure 5: Hidden information about education decisions: Explanatory power loss from predicting future outcomes using item-level vs aggregate test data

Notes: This figure shows improvements in predictive power over long-run outcomes from using items over aggregates. We measure gains in predictive power using the difference in out-of-sample R-squared between model (10) and model (11). We benchmark the difference by scaling the gains in predictive power by the R-squared of the model trained on aggregates. Thus, our results communicate how much additional predictive power we gain for decision making relative to a baseline of using only aggregates. To ensure that any gains in predictive power are purely measuring additional information in the items, rather than the ability of the items to better interact with the cohort variable and detect changes over time, we remove all time series variation from our fit metrics by calculating R-squared within grade-cohort (so holding c_{ig} in models 10 and 11 constant) and then averaging across cohorts for a given grade-level. “Class failure” refers to results from models where the outcome is an indicator for failing at least one class in the indicated grade. “Disciplinary action” refers to results from models where the outcome is an indicator that takes a value of one if the student had any disciplinary action in the subsequent year. “Highschool graduation” refers to results from models where the outcome is an indicator for graduating high school. “College attendance” refers to results from models where the outcome is an indicator for ever attending college. “Wages” refers to results from models where the outcome is earnings 5 years after on-time graduation; however, we caution that these results are estimated for a single cohort of students since we are only able to create the appropriate data linkages for that lone cohort.

(a) Predicted math class failure



(b) Predicted ELA class failure

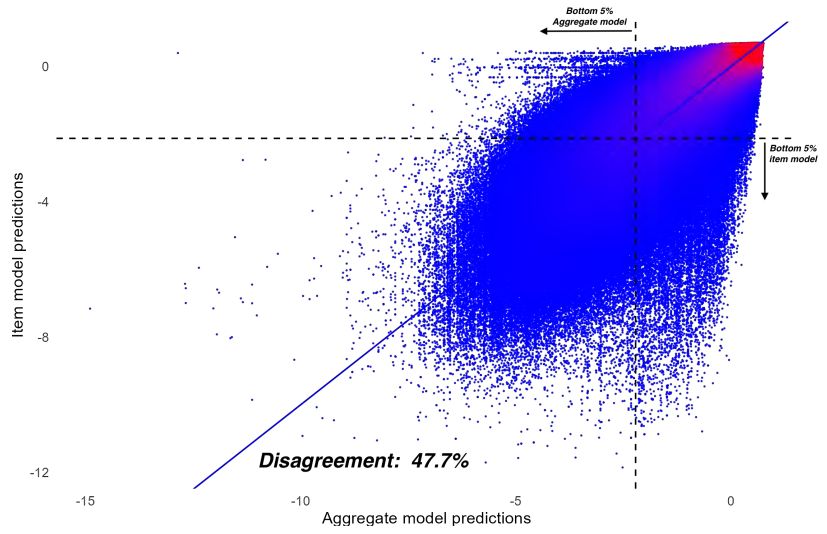
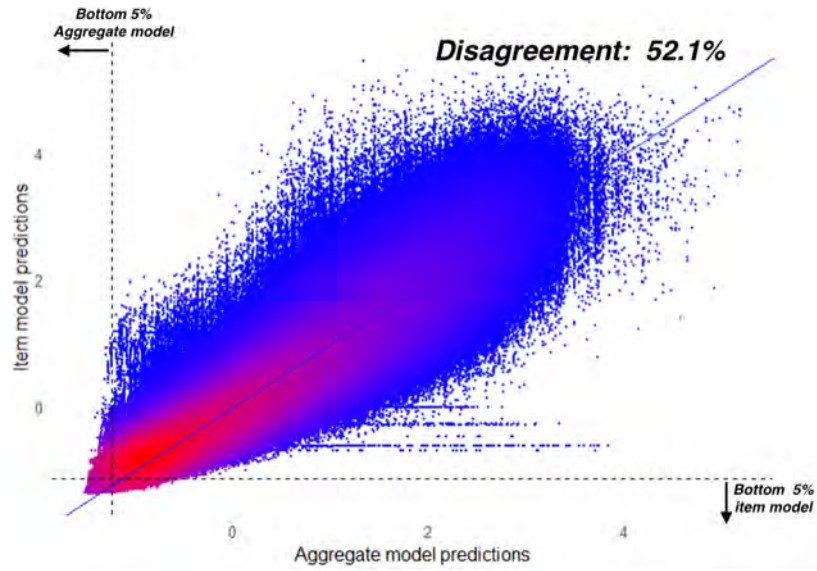


Figure 6: Disagreement between aggregate test score versus item-level test data regarding which students are struggling (Panels a–b).

(c) Predicted disciplinary action



(d) Predicted high school graduation

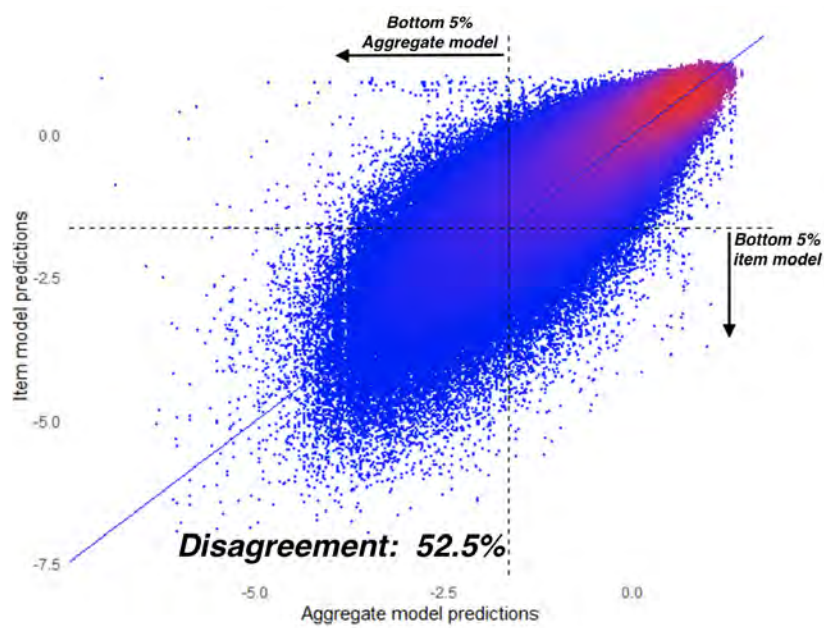


Figure 6: Disagreement between aggregate test score versus item-level test data regarding which students are struggling (Panels c–d).

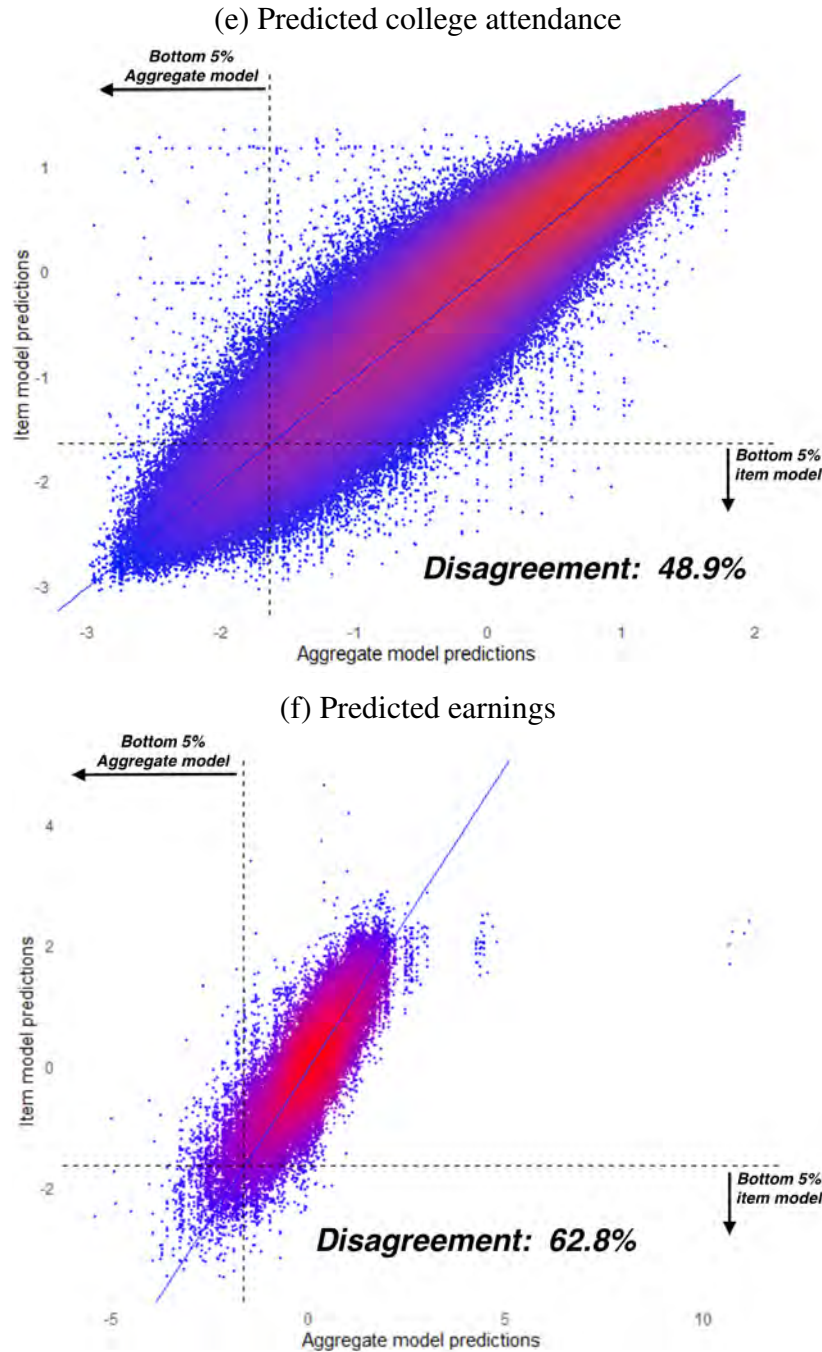
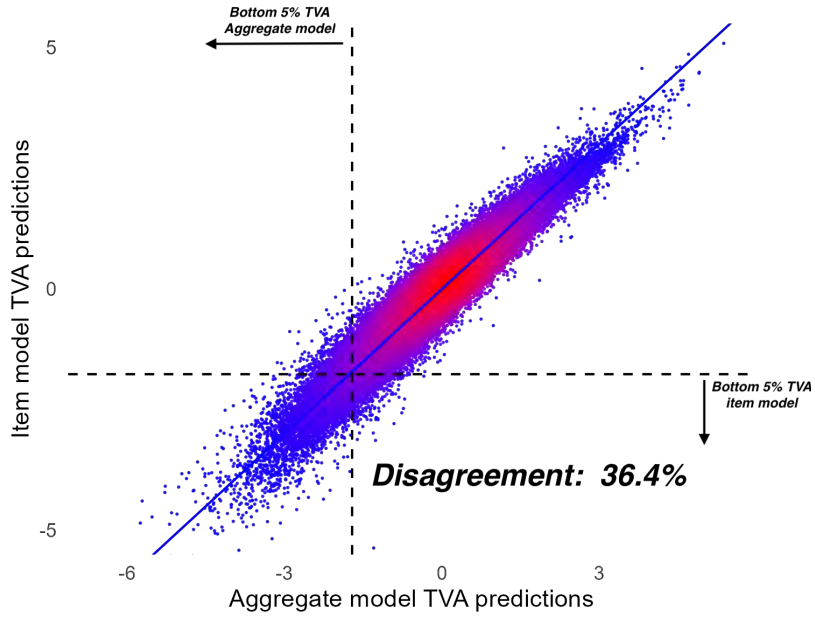


Figure 6: Disagreement between aggregate test score versus item-level test data regarding which students are struggling (Panels e–f).

Notes: This figure shows that items paint a very different picture than aggregates with respect to who is on track for success. Panel (a) plots the predicted values from the model that uses aggregate test data (X-axis) versus the item-level test data (Y-axis) to predict failing a math class. Panels (b) through (f) replicate the analysis for failing an ELA class, disciplinary action, highschool graduation, college attendance, and earnings. In all cases we standardize the predicted values to have mean zero, standard deviation 1 prior to plotting. Dashed lines denote the bottom 5% threshold for each axis. The solid line is the 45-degree line. The color gradient is a visual depiction of the density (red = high density, blue = low density).

(a) TVA predicted math class failure



(b) TVA predicted ELA class failure

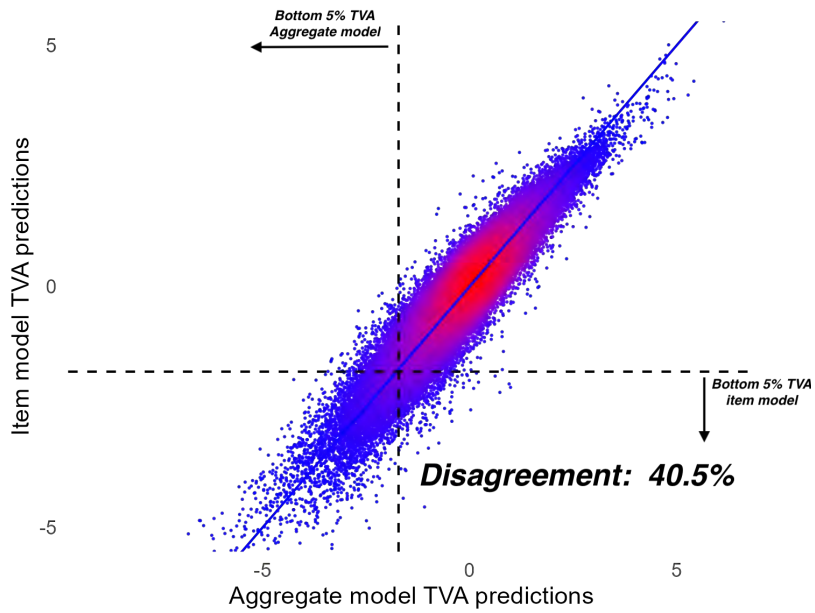


Figure 7: Disagreement between aggregate test score versus item-level test data regarding which teachers are struggling (Panels a–b).

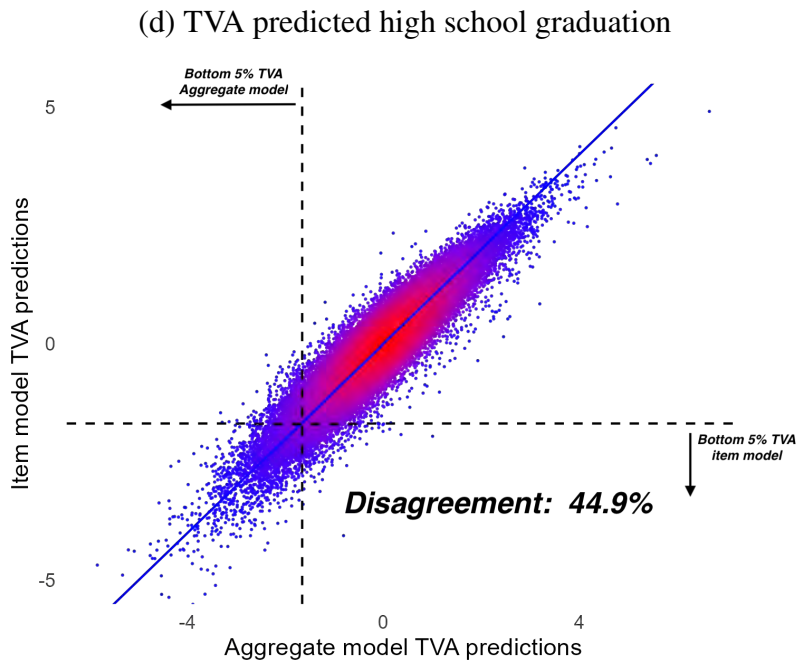
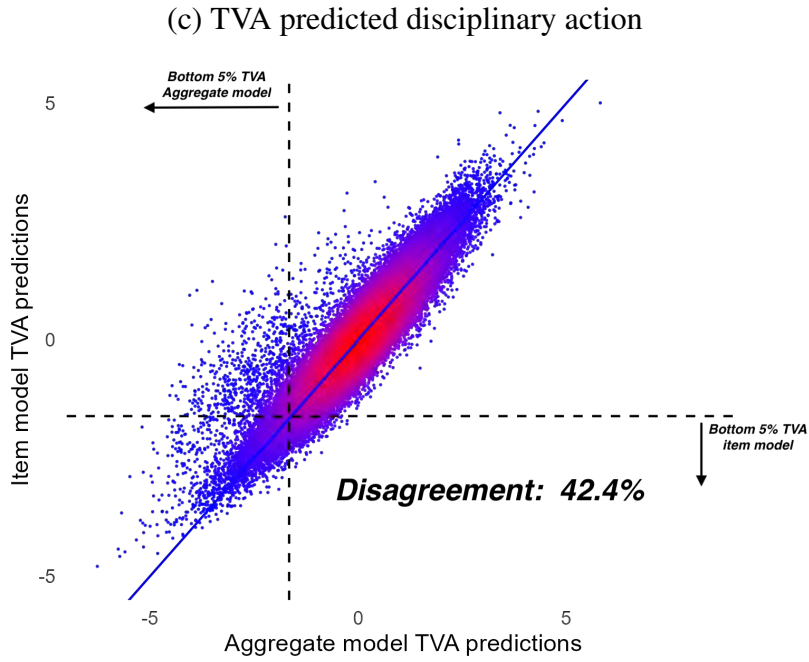


Figure 7: Disagreement between aggregate test score versus item-level test data regarding which teachers are struggling (Panels c–d).

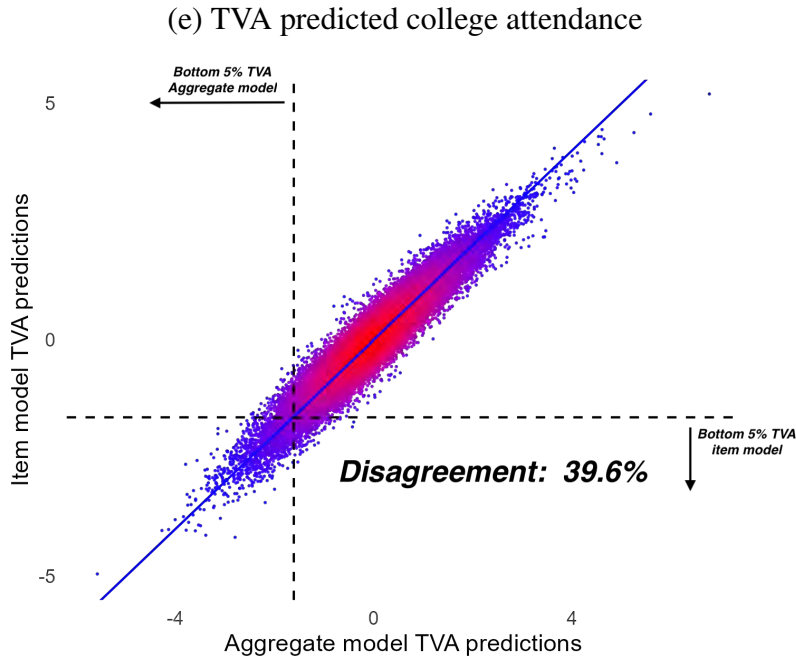


Figure 7: Disagreement between aggregate test score versus item-level test data regarding which teachers are struggling (Panel e).

Notes: This figure shows how the extra information in the test-score items leads us to rank-order teachers differently relative to aggregates. To generate this figure, we used the predicted values from models (10) and (11) as the left-hand side variable for the teacher value added model described in equation (8). The y-axis plots teacher value-add on predicted values formed using items. The x-axis plots teacher value-add on predicted values formed using aggregate scores. Panels (a) through (e) plot results for predicted math class failure, ELA class failure, disciplinary action, high school graduation, and college attendance. We cannot perform this exercise for earnings due to the limited data linkages for this outcome. Dashed lines denote the bottom 5% threshold for each axis. The solid line is the 45-degree line. The color gradient is a visual depiction of the density (red = high density, blue = low density).

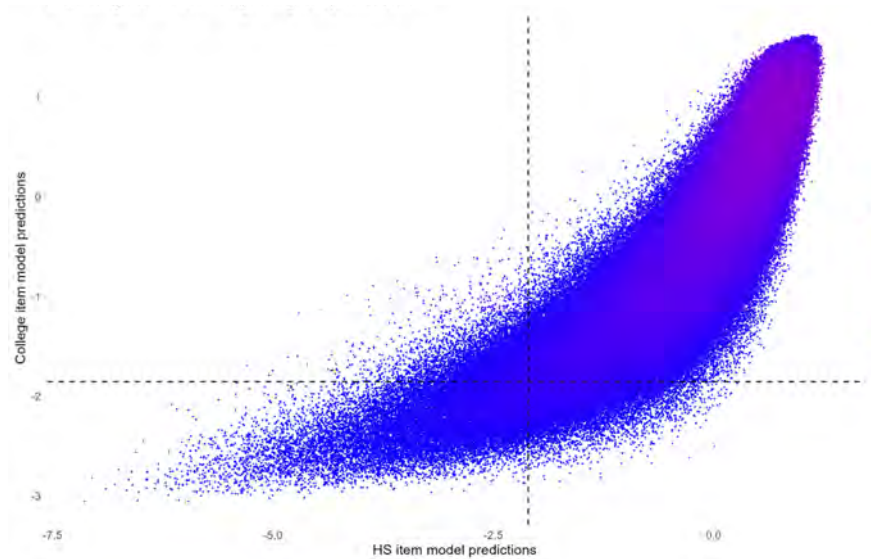


Figure 8: Plotting predicted high school graduation vs. college attendance from models that use item-level test data

Notes: This figure demonstrates that different test items carry different amounts of signal about different outcomes. In this figure, the x-axis denotes item-level predicted values from model (10) with high-school graduation as the left hand side variable. The y-axis denotes item-level predicted values from model (10) with college attendance as the left hand side variable. Thus each observation is a student, with the color gradient giving a visual depiction of the density (red = high density, blue = low density). Predicted values are standardized to have mean zero, standard deviation 1 prior to plotting. The wide dispersion in the data implies that some items carry different predictive power over different outcomes and hence there cannot be a single, optimal aggregator. Appendix figure A.6 presents comparable results for teachers.

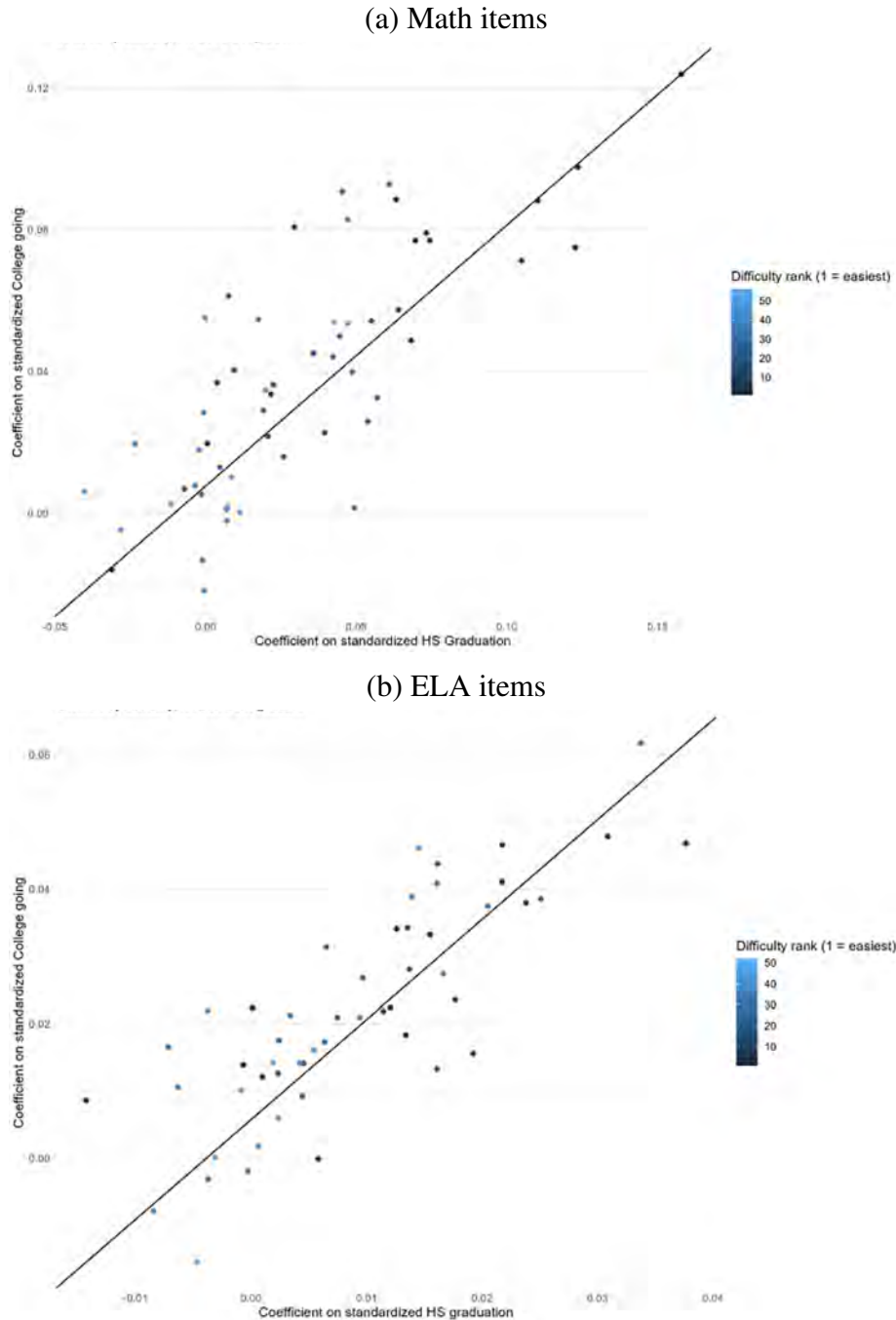
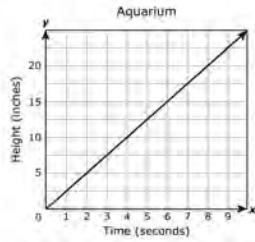


Figure 9: Individual test items have different signal for different outcomes – high school graduation vs. college attendance (conditional on overall ability)

Notes: This figure shows that different items contain different predictive power for long run outcomes even conditional on ability. Both panels of this figure plot item slopes from model 13 estimated with high school graduation as the outcome (x-axis) against item slopes from model 13 estimated with college attendance as the outcome (y-axis). The solid line is the line of best fit. Points are color coded by difficulty (i.e. the share of students that answered the item correctly). The sample for this figure is restricted to students who took the 8th grade test in 2016. Importantly, because these coefficients are estimated with every student in Texas who took the test in this grade and year (57,594 in math and 57,595 in ELA), the standard errors on the individual slope estimates are small (ranging from 0.003 to 0.015 in standard deviation units), so we omit measures of sampling uncertainty for visual clarity. Panel (a) gives results for math. Panel (b) gives results for ELA.

(a) Question 5 in 2016

5 An aquarium is being filled with water. The graph shows the height of the water over time as the aquarium is being filled.



Which statement best describes the rate of change for this situation?

- A The height of the water increases 20 inches per second.
- B The height of the water increases 1 inch per second.
- C The height of the water increases 5 inches per second.
- D The height of the water increases 2.5 inches per second.

(b) Question 5 in 2017

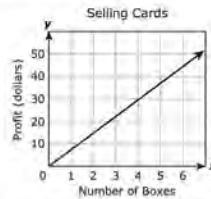
5 Triangle MNP is graphed on a coordinate grid with vertices at $M(-3, -6)$, $N(0, 3)$ and $P(6, -3)$. Triangle MNP is dilated by a scale factor of u with the origin as the center of dilation to create triangle $M'N'P'$.

Which ordered pair represents the coordinates of the vertex P' ?

- A $(6 + u, -3 + u)$
- B $(\frac{6}{u}, -\frac{3}{u})$
- C $(6 + \frac{1}{u}, -3 + \frac{1}{u})$
- D $(6u, -3u)$

(c) Question 39 in 2017

39 Emily sells greeting cards. The graph models the linear relationship between the number of boxes of cards she sells and her profit.



Which of these best describes the profit Emily makes from selling these cards?

- A \$7.50 per box
- B \$10.00 per box
- C \$4.00 per 30 boxes
- D \$3.00 per 4 boxes

Figure 10: Select items from the Grade 8 Exam in 2016 and 2017

Notes: This figure highlights the challenge of leveraging item level variation for practical decision making: without additional information, it is difficult to “link” items across test years in a way that permits the extrapolation of long-run relationships forward in time so that they can be leveraged for the decisions we want to make about students and teachers today. Panel (a) displays question 5 on the grade 8 test administered in year 2016. Panel (b) displays question 5 from the grade 8 test administered in 2017. Despite the fact that both of these questions are labeled as “item 5,” they clearly capture very different mathematical concepts. However, panel (c) displays question 39 from the 2017 test, which is nearly identical in structure and content to question 5 from the 2016 test. Thus, to make item level variation practically useful, it is critical to develop ways of linking “similar” items across test years.

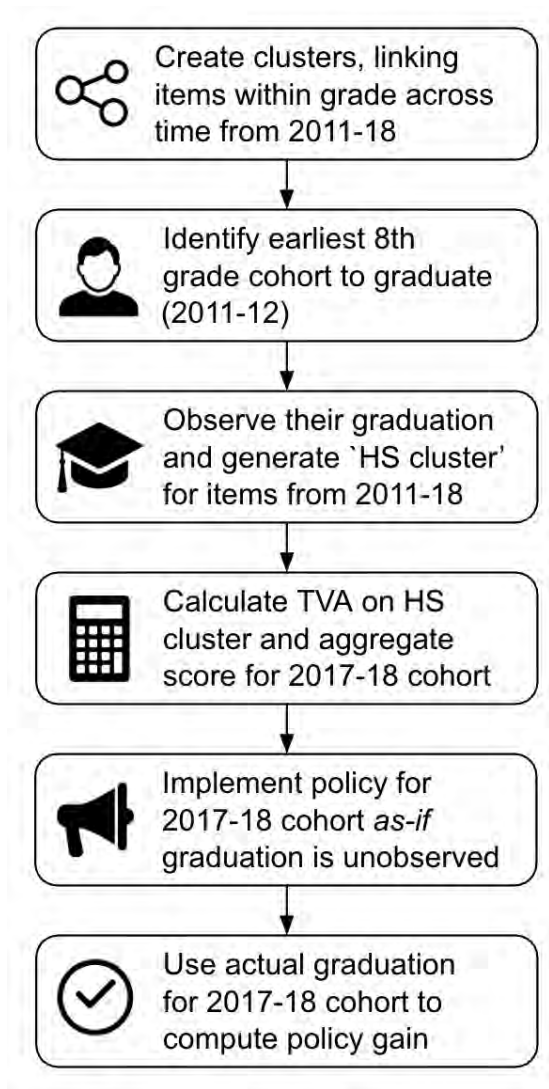
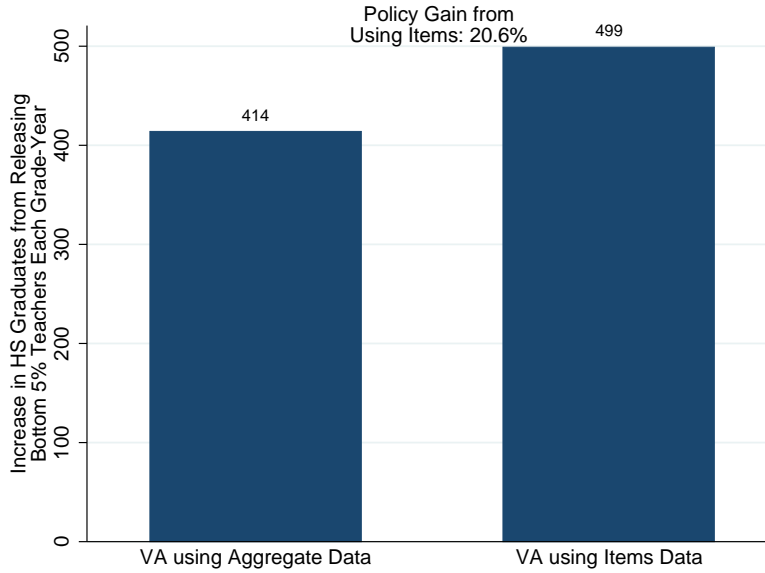


Figure 11: Schematic of evaluating a feasible implementation of item approach to decision-making

Notes: This diagram highlights our approach to evaluating whether an item-based approach will have a large practical impact if implemented feasibly, whereby “feasibly” we mean in the sense that a policy maker needs to make decisions today using items (or aggregates) for students for whom they cannot directly observe future outcomes like high school graduation. To accomplish this, we first use the algorithm from section 7.1 to link items within-grade, over time from 2011-2018. Next, we identify the earliest cohorts of students in our data to actually graduate. Next, we use their observed graduation to identify the cluster of items most highly correlated with high school graduation. Then, we calculate teacher value-added on the high school cluster and the average score, and use those estimates to identify which teachers fall into the bottom 5% of the distribution for the 2 cohorts of 8th graders, and one cohort of 7th graders, where we can observe their high school graduation (but which were not used to identify the high school cluster). Finally, we use observed high school graduation for these cohorts to calculate the implied change in high school graduation rates if we used teacher value-added on the high-school cluster instead of teacher value-added on the average score to implement a hypothetical policy that replaced the bottom 5% of teachers with an average one.

(a) Average gain in high school graduates per grade-year



(b) Marginal value of public funds

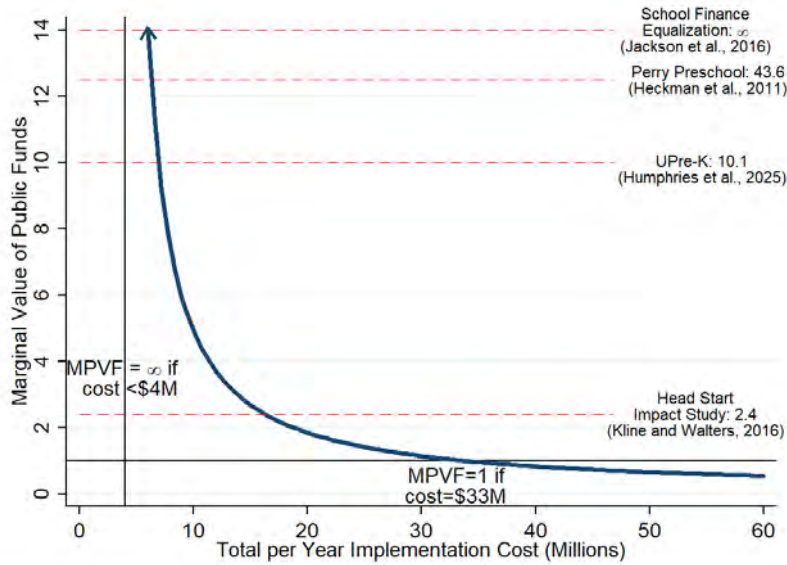


Figure 12: Feasible Gains: Items vs. Aggregates

Notes: Figure 12(a) presents the average gain per cohort-year, measured in terms of newly created graduates, from implementing the bottom 5% policy with traditional value-add and our feasible, item based approach. Relative to the aggregate, the item based approach generates a 20.6% improvement. Figure 12(b) displays the marginal value of public funds (MPVF) for the additional high school graduates gained by using our feasible, item-based implementation of the policy. The $MPVF = \frac{(1-\tau) \sum_i \Delta Earnings_i}{Cost - \tau \sum_i \Delta Earnings_i}$ where τ is the marginal tax rate and $\sum_i \Delta Earnings_i$ represents the change in present-value monetary earnings from being a high school graduate relative to a high school dropout. For the earnings estimates, we assume the average employed high school dropouts earns \$38,500, two-thirds of high school dropouts are employed, a 10% return to high school graduation (Card, 1999), a discount rate of 3%, a federal marginal tax rate of 12%, and that testing runs for six grades (grades 3-8). Red dots indicate MPVF calculations from four other well-known programs: Head Start Impact Study (Kline and Walters, 2016), Perry Preschool (Heckman et al., 2010), school finance equalization (Jackson et al., 2015), and universal pre-Kindergarten Humphries et al. (2024). MPVFs for Heckman et al. (2010); Jackson et al. (2015); Kline and Walters (2016) are taken from Hendren and Sprung-Keyser (2020).

Table 1: Summary Statistics

	Full Sample ^a (Grade 3-8) (1)	Test-Taking Sample ^b (Grade 3-8) (2)	Value-Added Sample ^c (Grade 4-8) (3)
<i>Test Statistics</i>			
# of Items Math Test	-	45.3	46.3
% Correct on Math Test	-	61.6	60.0
# of Items ELA Test	-	43.4	44.6
% Correct on ELA Test	-	66.5	66.3
<i>Demographics</i>			
% Hispanic	51.8	51.0	51.5
% White	29.5	29.9	30.0
% Black	12.7	13.2	13.1
% Asian	3.6	3.4	2.9
% Disadvantaged	60.7	60.6	61.2
% Limited English Proficient	16.7	14.7	13.0
% Special Education	9.0	6.8	5.6
% Gifted	9.9	8.9	7.3
Class Size	-	-	22.0
<i>Outcomes</i>			
Graduated High School (%) ^d	79.1	79.6	81.1
Entered College (%) ^e	61.6	62.3	61.5
Wages (Annual) ^f	18,061.8	17,623.9	18,038.0
Next Grade Disciplinary Infraction (%)	16.5	16.5	17.1
Failed Class (%) ^g	4.2	4.2	4.2
Observations (Student-item-year)	-	1,313,711,814	823,391,476
Observations (Student-year)	16,928,123	14,817,051	9,066,474
# of Students	4,912,003	4,650,274	3,649,525
# of Teachers	-	-	77,346
# of Schools	7,529	7,433	6,899

^a Data coverage: All students in grades 3-8 from 2011-12 through 2018-19 with valid math tests.

^b Data coverage: Same as column (1), but restricted to student-year observations with valid test scores for both math and reading. A valid test refers to the standard version of the STAAR test where all items were not left blank.

^c Data coverage: Same as column (2), but restricted to students who: (i) have valid lagged test scores (dropping third grade), (ii) can be uniquely matched to a math teacher, and (iii) are in a class with between 5 and 200 students.

^d To ensure students have reached the graduation age, we restrict our sample for each grade to the following years: grade 3 up to 2011-12, grade 4 up to 2012-13, grade 5 up to 2013-14, grade 6 up to 2014-15, grade 7 up to 2015-16, and grade 8 up to 2016-17.

^e To ensure students have reached college-going age, we restrict our sample for each grade to the following years: grade 4 up to 2011-12, grade 5 up to 2012-13, grade 6 up to 2013-14, grade 7 up to 2014-15, and grade 8 up to 2015-16.

^f Average annual wages (including zero wages) five years after expected high school graduation. 28.6% (45,723/159,633) of our wage sample have zero wages (defined as earning less than \$500). Only covers the 2011-12 grade 8 cohort.

^g Indicates failing a class in mathematics, ELA, reading, social studies, or science.

Table 2: Share of information lost in aggregation (students)

	Information Loss					F-stat	P-Value	Obs
	2016	2017	2018	2019	16-19			
Grade 4	20.5%	14.2%	16.5%	17.5%	17.1%	1.5	0.00	37,116,226
Grade 5	15.0%	16.6%	20.5%	20.7%	18.2%	1.6	0.00	40,213,906
Grade 6	9.0%	9.6%	12.7%	14.5%	11.4%	1.7	0.00	38,898,566
Grade 7	14.1%	10.2%	10.9%	12.2%	11.9%	1.7	0.00	38,827,074
Grade 8	15.5%	12.4%	16.4%	17.9%	15.6%	1.8	0.00	36,511,118

Notes: This table presents estimates of the information about students that is lost in aggregation. Each row denotes a separate grade of enrollment. The first four columns denote AY 2016-2019, while the fifth column (labeled 16-19) summarizes average information loss over all years within a given grade. The measure of information loss is an estimate of $\frac{\text{var}(\gamma_{ic} - \gamma_i)}{\text{var}(\gamma_{ic})}$. The F-stat and associated p-value are from a formal test that there is no information lost by assuming a single latent factor (i.e. that $\gamma_{ic} = \gamma_i \forall q$). Observations represent the number of student-item-years used in the data to generate each row. Finally, we note that these results are likely to be a lower bound, since we are forced to rely on imperfect categories rather than directly estimating the relevant student-by-item fixed effects.

Table 3: Share of information lost in aggregation (teachers)

	Information Loss					F-stat	P-Value	Obs
	2016	2017	2018	2019	16-19			
Grade 4	66.2%	57.6%	63.7%	62.8%	62.6%	1.8	0.00	37,120,964
Grade 5	62.8%	64.7%	61.9%	60.4%	62.4%	1.9	0.00	40,352,816
Grade 6	67.0%	60.7%	61.5%	63.1%	63.0%	2.3	0.00	38,930,276
Grade 7	71.9%	68.8%	71.6%	69.8%	70.2%	2.2	0.00	38,862,708
Grade 8	74.1%	69.3%	68.6%	69.6%	70.2%	2.8	0.00	36,641,290

Notes: This table presents estimates of the information about teachers that is lost in aggregation. Each row denotes a separate grade of enrollment. The first four columns denote AY 2016-2019, while the fifth column (labeled 16-19) summarizes average information loss over all years within a given grade. The measure of information loss is an estimate of $\frac{\text{var}(\delta_{jq} - \delta_j)}{\text{var}(\delta_{jq})}$. The F-stat and associated p-value are from a formal test that there is no information lost by assuming a single latent factor (i.e. that $\delta_{jq} = \delta_j \forall q$). Observations represent the number of teacher-student-item-years used in the data to generate each row.

Table 4: Disagreement at the bottom of the distribution (students)

	2012	2013	2014	2015	2016	2017	2018	2019	12-17	P-Value	Obs.
<i>Panel A: Predicted graduation</i>											
Grade 4	55.3%	48.9%							48.4%	0.00	114,761
Grade 5	47.8%	51.7%	51.6%						50.0%	0.00	177,415
Grade 6	55.6%	58.7%	57.4%	55.0%					56.6%	0.00	230,523
Grade 7	54.7%	55.4%	54.2%	53.4%	58.8%				55.1%	0.00	279,675
Grade 8	53.6%	51.2%	53.6%	47.2%	45.2%	50.8%			50.1%	0.00	320,374
<i>Panel B: Predicted college</i>											
Grade 4	46.6%								46.6%	0.00	57,978
Grade 5	47.1%	50.2%							49.0%	0.00	118,360
Grade 6	50.5%	55.7%	51.8%						51.2%	0.00	172,082
Grade 7	51.2%	52.7%	51.1%	48.3%					50.0%	0.00	221,763
Grade 8	49.4%	45.8%	51.1%	45.8%	43.2%				47.0%	0.00	258,497
<i>Panel C: Predicted wages</i>											
Grade 8	62.8%								62.8%	0.00	39,284
<i>Panel D: Predicted disciplinary infraction</i>											
Test grade 4	50.5%	50.2%	47.6%	44.3%	48.5%	55.9%	55.9%		50.4%	0.00	403,639
Test grade 5	38.1%	48.4%	43.9%	43.7%	43.0%	40.8%	40.2%		44.8%	0.00	409,265
Test grade 6	41.1%	47.4%	45.9%	43.0%	46.5%	45.6%	47.8%		49.1%	0.00	403,907
Test grade 7	57.3%	55.3%	57.9%	57.7%	58.7%	69.6%	58.5%		64.4%	0.00	391,744
Test grade 8	48.5%	49.5%	48.1%	79.0%	57.7%	57.3%	58.1%		61.2%	0.00	376,267
<i>Panel E: Predicted pass class</i>											
Grade 4	45.4%	43.8%	47.1%	48.0%	45.7%	47.8%	50.6%	51.7%	46.6%	0.00	2,078,976
Grade 5	45.9%	45.6%	42.5%	44.8%	44.2%	45.4%	43.8%	39.8%	41.8%	0.00	2,159,728
Grade 6	53.4%	51.8%	48.8%	53.3%	55.5%	56.9%	55.2%	54.9%	53.5%	0.00	1,251,104
Grade 7	52.6%	52.6%	49.3%	52.7%	54.4%	57.1%	51.1%	53.6%	52.9%	0.00	1,967,294
Grade 8	54.8%	52.5%	56.1%	51.9%	54.5%	55.5%	55.4%	55.9%	52.9%	0.00	1,723,489

Notes: This table shows that there is substantial disagreement between item-level and aggregate-level predictions regarding which students are not on track for success with respect to various outcomes as measured by falling into the bottom 5% of the distribution of predicted values from models (11) and (10). Specifically, we calculate the disagreement share by first counting the total number of students that fall into the bottom 5% of the distribution of predictions made using either items or aggregates (but not both), and then divide by the total number of students that falls into the bottom five percent of at least one of the two distributions. We follow the approach from [Petek and Pope \(2023\)](#); [Rose et al. \(2022\)](#) by standardizing the predictions to each have mean 0 and standard deviation 1, and then counting a student in the bottom 5% of the relevant distribution if the value falls below -1.645. This approach will be exactly correct under the assumption that the predicted values are normally distributed, and has the advantage that it simplifies inference considerably relative to defining the bottom 5% using the corresponding empirical quantile. However, we find similar point estimates using a threshold that is based on the empirical 5th quantiles (see appendix [A.11](#)). P-values test the null of no-disagreement in the bottom 5% across all cohorts in a given row using a bootstrap procedure that accounts for estimation error in the random-forest fitting stage (see appendix [A.12](#) for more detail).

Table 5: Disagreement at the bottom of the distribution (teachers)

	2012	2013	2014	2015	2016	2017	2018	2019	12-17	P-Value	Obs
<i>Teacher value-added on:</i>											
Predicted graduation	44.3%	42.1%	48.6%	41.3%	43.8%	40.0%			44.9%	0	68,933
Predicted college attendance	41.6%	39.1%	40.6%	41.6%	34.6%				39.6%	0	42,007
Predicted disciplinary infraction	41.2%	42.0%	41.7%	39.6%	42.6%	42.8%	44.8%		42.1%	0	148,105
Predicted pass/fail class	51.0%	51.4%	53.0%	50.6%	51.6%	50.5%	52.0%	52.1%	51.3%	0	608,000

Notes: This table shows that there is substantial disagreement about which teachers are struggling (i.e. falling into the bottom 5%) using value-add estimates formed from items that are connected to long-run outcomes relative to teacher value-add estimates formed using aggregate scores. Specifically, we calculate the disagreement share by first counting the total number of teachers that fall into the bottom 5% of the distribution of value-add formed using either items or aggregates (but not both), and then divide by the total number of teachers that fall into the bottom five percent of at least one of the two distributions. We follow the approach from [Petek and Pope \(2023\)](#); [Rose et al. \(2022\)](#) by standardizing the value-add estimates to each have mean 0 and standard deviation 1, and then counting a teacher in the bottom 5% of the relevant distribution if the value-add falls below -1.645. This approach will be exactly correct under the assumption that the value-added estimates are normally distributed, and has the advantage that it simplifies inference considerably relative to defining the bottom 5% using the corresponding empirical quantile. However, we find similar point estimates using a threshold that is based on the empirical 5th quantiles (see appendix [A.11](#)). P-values test the null of no-disagreement in the bottom 5% across all cohorts in a given row using a bootstrap procedure that accounts for estimation error in the random-forest fitting stage (see appendix [A.12](#) for more detail).

Online Appendix for “Do Test Scores Misrepresent Test Results? An Item-by-Item Analysis”

By Jesse Bruhn, Michael Gilraine, Jens Ludwig, and Sendhil Mullainathan

Online Appendix Table of Contents

A	Additional details of empirical exercises, supplemental results, and robustness	2
A.1	Details of categorization algorithm and sensitivity to alternative methods	2
A.2	Placebo study simulation details	3
A.3	Estimating information loss	3
A.4	Robustness of teacher value-added rank swapping to alternative estimation procedures	4
A.5	Value-Added Controls	4
A.6	Jack-knife, empirical Bayes procedure for item-level teacher value-add	5
A.7	Equivalence of traditional TVA and item-level TVA under homogeneity	6
A.8	Measurement error corrected rank-correlations	6
A.9	Robustness of long run outcome results	7
A.10	Split sample IV tests	7
A.11	Robustness to using empirical quantiles	8
A.12	Details of bootstrap procedure	9
A.13	Additional tables and figures	10
B	Proof of Proposition 1	11

A Additional details of empirical exercises, supplemental results, and robustness

A.1 Details of categorization algorithm and sensitivity to alternative methods

In Section 5, we show that both students and teachers exhibit comparative advantage at the item-by-item level that is missed when focusing on the aggregates. In this appendix, we provide some additional details, as well as some robustness checks.

As discussed in the main text, our ideal way of quantifying the item comparative advantage at the student level would be to estimate student-by-question interaction terms; however, these are not identified, since every student answers each question only once. Thus, we test a related implication: if an individual student’s performance is, in expectation, identical on every item after conditioning on question difficulty, then the individual student’s performance should also be identical on every sub-set of the test.

Our main results are based on categories found via a k-means clustering algorithm. Intuitively, this will cluster questions within a grade level according to whether students tend to get them correct / incorrect together.⁴⁷

More precisely, we find the categories using the following procedure:

1. Find every student that took a valid test in a given grade in a given year (e.g. the 8th grade test in 2012).
2. Reshape the data so that each row is a student, each column is an item from that test, and each cell is a binary variable that takes a value of 1 if the student correctly answered the item corresponding to that column.
3. Partition the items into five categories in order to minimize within-partition distance between the average (across all students) performance within that partition and individual student performance.⁴⁸
4. Repeat steps 1-3 for every other grade and year.

However, our results are not sensitive to this decision. In appendix figure A.1, we show that similar results obtain if we use content categories provided by the state of Texas, and if we use categories derived from the item linking algorithm developed in Section 7.

[Figure A.1 about here.]

⁴⁷Note: the algorithm we use to find these categories is distinct from the one we develop in section 7; however, we show later on in this appendix that we obtain similar results using the categories from section 7 as well.

⁴⁸In practice, we execute this k-means clustering using the off-the-shelf implementation in base R (the *kmeans* function) with the “nstarts” parameter set to 25. We chose five categories to match the number of content categories that the state of Texas attaches to the items.

A.2 Placebo study simulation details

In this section, we describe the details of the placebo study simulations in figures 3 and 4 that are meant to assess the potential for rank-swapping to emerge as a result of noise alone.⁴⁹

Student level placebo study. To conduct the placebo study in figure 3, we implement the following procedure:

1. Estimate model 5, and denote the standard error for $\hat{\gamma}_{ic(q)}$ as $SE_{\hat{\gamma}_{ic(q)}}$
2. Re-estimate model 5, but imposing the null hypothesis that $\gamma_i = \gamma_{ic(q)} \forall c$. This yields an estimate of average performance across all test items for each student: $\hat{\gamma}_i$.
3. Draw placebo values for the student-category interactions under the null hypothesis according to the following distribution: $\tilde{\gamma}_{ic(q)} \sim N(\hat{\gamma}_i, SE_{\hat{\gamma}_{ic(q)}})$. Observe that $\tilde{\gamma}_{ic(q)}$ is just average performance, but perturbed by random noise from the sampling distribution of the appropriate category-by-student level interaction term.
4. Plot $\hat{\gamma}_i$ in the leftmost column (average performance) and plot $\tilde{\gamma}_{ic(q)}$ under the appropriate categories in the columns to the right (randomly perturbed average performance).

Teacher level placebo study. The teacher level placebo study on display in figure 4 is constructed identically to the algorithm described above for students, except that we use the relevant teacher level equations (i.e. equation 8) and corresponding teacher level parameters (i.e. $\delta_{j(i)q}$ and $\delta_{j(i)}$) in place of the student level counterparts (i.e. $\gamma_{ic(q)}$ and γ_i).

A.3 Estimating information loss

For students, we estimate the share of information lost by comparing: (1) the total amount of variation in performance we can explain while allowing student ability to be different across categories, to (2) the amount of variation that we can explain under the null of a single latent factor. Specifically, we do this via an ANOVA type procedure that involves estimating three fixed effect regressions:

$$d_{iq} = \omega_q + \eta_{iq} \tag{15}$$

$$d_{iq} = \omega_q + \gamma_{ic(q)} + u_{iq} \tag{16}$$

$$d_{iq} = \omega_q + \gamma_i + e_{iq} \tag{17}$$

Where ω_q is a question fixed effect; $\gamma_{ic(q)}$ are student-by-category fixed effects; and γ_i is a student fixed effect.

Observe that the residual variance of equation 15 is: $var(\eta_{iq}) = var(\gamma_{ic(q)}) + var(u_{iq})$. Thus the difference in residual variance between equations 15 and 16 identifies the total amount of variation in student performance explained when student ability is allowed to vary across categories:

$$var(\eta_{iq}) - var(u_{iq}) = var(\gamma_{ic(q)}) = \text{total variation from multi-dimensional ability.} \tag{18}$$

Similarly, observe that $var(e_{iq}) = var(\gamma_{ic(q)} - \gamma_i) + var(u_{iq})$. Thus, the difference in residual

⁴⁹This is also the algorithm we use for the corresponding robustness checks in appendix figure A.1.

variance between equations 16 and 17 identifies the variation we overlook when treating ability as homogeneous:

$$\text{var}(e_{iq}) - \text{var}(u_{iq}) = \text{var}(\gamma_{ic(q)} - \gamma_i) = \text{variation lost assuming constant ability.} \quad (19)$$

Thus, we can compare these two variance terms to identify the share of the systematic component of test performance which is lost by incorrectly assuming student ability is constant across categories:

$$\text{Information loss (\%)} = \frac{\text{var}(e_{iq}) - \text{var}(u_{iq})}{\text{var}(\eta_{iq}) - \text{var}(u_{iq})} = \frac{\text{var}(\gamma_{ic(q)} - \gamma_i)}{\text{var}(\gamma_{ic(q)})} \quad (20)$$

In practice, we estimate the residual variances necessary to calculate information loss using estimates of mean-square error that apply the standard degrees of freedom correction. For example: $\hat{\text{var}}(\eta_{iq}) = \frac{\sum(d_{iq} - \hat{\omega}_{iq})^2}{N-Q}$ where N is the number of student-by-question pairs, and Q is the number of items. Applying the degrees of freedom correction ensures that the resulting estimates of mean squared error are unbiased: $\mathbb{E}(\hat{\text{var}}(\eta_{iq})) = \text{var}(\eta_{iq})$ (Wooldridge, 2010).

For teachers, the procedure is identical, except that we compare the residual variance in student performance from equation 8 to similar models that assume homogeneous teacher effects: ($\delta_j = \delta_{jq} \forall q$).

A.4 Robustness of teacher value-added rank swapping to alternative estimation procedures

In section 5.2, our baseline model uses item-level teacher-value added estimates that (1) are adjusted for a classroom shock using an adapted version of the jack-knife leave-year-out procedure from Chetty et al. (2014a) (see appendix A.6 for details); and (2) only control for baseline ability using the prior year’s average test score. In figure A.2, we show that the results in the main text are robust to this decision. In panels (a) and (b), we show that we obtain similar results if we use the unadjusted fixed effects implied by equation 8. In panels (c) and (d), we show that we obtain similar results if we use the question-by-lagged score interactions implied by equation 7.

[Figure A.2 about here.]

A.5 Value-Added Controls

Following Chetty et al. (2014a), we control for the following student-level controls: (i) lagged test scores using a cubic polynomial in prior-year scores in math and reading, interacted with grade dummies, (ii) demographics, including: ethnicity (six ethnic groups), economically disadvantaged status (four groups), gender, limited English status (three groups), special education status, and gifted status. We also include the following class- and school-grade level controls: (i) cubics in class and school-grade means of prior-year test scores in math and reading interacted with grade dummies, (ii) class and school-grade means of all the demographic covariates, (iii) class size, (iv) grade dummies, and (v) year dummies.

A.6 Jack-knife, empirical Bayes procedure for item-level teacher value-add

Recall that, in order to generate estimates of teacher value-added on average test scores that account for the classroom shock component of teacher value-add variance, [Chetty et al. \(2014a\)](#) develop a jack-knife empirical-Bayes procedure that effectively generates estimates of value-add for year t by predicting it with teacher value-add estimates from all years other than t . To adapt this idea to the item level, we use teacher value-add on all *items* in years other than t to predict value-add on item q in year t .

Formally, let \hat{tva}_{j_tq} denote the estimated value-add of teacher j on question q in year t .⁵⁰ Observe that we can write this estimate as:

$$\hat{tva}_{j_tq} = tva_{j_tq} + e_{j_tq} \quad (21)$$

Which just says that our estimate is equal to actual teacher value-add (tva_{j_tq}) plus noise (e_{j_tq}).

Now, decompose the sampling variation as follows:

$$e_{j_tq} = \eta_{qc(t,j)} + \epsilon_{j_tq} \quad (22)$$

Where $\eta_{qc(t,j)}$ is a year and question specific, classroom component to the item level teacher value-add variance that does not disappear asymptotically (class sizes are generally capped at 15-30 students in a given year); and ϵ_{j_tq} is residual, mean zero idiosyncratic variation at the teacher-question-year level.

To eliminate the classroom shock, we take teacher value-add on items estimated in years other than t , and estimate regressions analogous to equation (6) in [Chetty et al. \(2014a\)](#):

$$\hat{tva}_{j_tq} = \sum_{\tau \in T_{-t}} \sum_{q' \in Q_{\tau}} \psi_{\tau q'} \hat{tva}_{j_{\tau}q'} + u_{j_tq} \quad (23)$$

Where T_{-t} is that set of all years of data other than t ; Q_{τ} is the set of all items given in year τ ; and u_{j_tq} is a residual.

Since the classroom component $\eta_{qc(t,j)}$ is, by definition, specific to that classroom (and hence does not persist across time), we know that:

$$\text{cov}(\eta_{qc(t,j)}, \eta_{q'c(t',j)}) = 0 \quad \forall q, q' \quad \text{and} \quad \forall t' \neq t \quad (24)$$

⁵⁰Note that in the main text, for simplicity, we suppress the t subscript since we can view it as encoded in the question index q (e.g. question 1 on the 2016 test would have a distinct index from question 1 on the 2017 test, etc.); however, it will be useful to make the dependence on time explicit in the notation here.

Thus, we can form estimates of item-level teacher value-added that are free from the classroom shock by using the estimated parameters from 23 to predict tva_{jtq} similar to the standard leave-year-out jackknife empirical-Bayes estimates commonly deployed for average scores.

For computational tractability, we compute each of our classroom shock adjusted item-level TVA estimates using three years of data – the year prior to the focal test, the year of the focal test, and the year after the focal test. So for example, to estimate item-level TVA on questions contained in the 4th grade test in 2015, we would use item-level TVA estimates from the 4th grade tests in 2016 and 2014 to predict item-level TVA in 2015. In practice, this necessitates dropping teachers who do not have three consecutive years of item data at the relevant grade level. However, this restriction is not consequential – we find similar results using the “unadjusted” teacher-by-item fixed effects for the full sample of teachers (see Appendix figure A.2).

A.7 Equivalence of traditional TVA and item-level TVA under homogeneity

To see that traditional teacher value-add is, up to a normalization, equivalent to item-level TVA under homogeneity, observe that if we impose homogeneous teacher effects, then equation 8 becomes:

$$d_{iq} = \omega_q + \delta_{j(i)} + \pi X_i + u_{iq} \quad (25)$$

Now, take the average of 8 with respect to $q \in T$ to get a traditional teacher value-add model:

$$Y_i = \omega + \delta_{j(i)} + \pi X_i + e_i \quad (26)$$

Where Y_i is the average test score $|T|^{-1} \sum_{q \in T} d_{iq}$; $\omega = |T|^{-1} \sum_{q \in T} \omega_q$ is the average item fixed effect; and $e_{iq} = |T|^{-1} \sum_{q \in T} u_{iq}$.

Note that, because equation 25 is a projection, we have that $\mathbb{E}(u_{iq}(\omega_q + \delta_{j(i)} + \pi X_i)) = 0$. It follows immediately that $\mathbb{E}(e_i(\omega + \delta_{j(i)} + \pi X_i)) = 0$ and hence equation 26 also identifies $\delta_{j(i)}$.

Thus we have shown that, up to the normalization typically imposed on test scores prior to estimation (i.e. that they be mean zero, standard deviation one) our item level approach is, under homogeneity, equivalent to the typical approach using average scores.

A.8 Measurement error corrected rank-correlations

Kitagawa et al. (2018) provides a method for constructing bias-corrected estimators of rank correlations in the presence of measurement error using a small error variance approximation. However, the method requires that the variance of the measurement error is known or assumed. Thus, we ask how our substantive conclusion that the rank correlation between item-level TVA and TVA on the aggregate score does not equal one varies with the overall level of measurement error.

More specifically, in figure A.3, we plot the assumed proportion of observed variance attributable to measurement error (x-axis) against the number of item-level value-added estimates where we would fail to reject that the Spearman rank correlation equals one using the Kitagawa et al. (2018) bias-correction method.

We find that there are items exhibiting a statistically significant deviation from one even if we assume that over 50% of the variation in item-level teacher value-added is due to measurement error. Thus, our conclusions are robust to a large degree of measurement error in the teacher-level estimates.

[Figure A.3 about here.]

A.9 Robustness of long run outcome results

In section 6.1 of the main text, we show that there are large information gains to using the exact answers chosen by students on individual test items in order to predict various short, medium, and long-run outcomes. In this section, we show that this finding is robust to both: (1) alternative ways of using the items to predict aggregate achievement; and (2) alternative ways of representing aggregate achievement. Panel (a) of appendix figure A.4 is identical to figure 5 from the main text, except that the underlying random forest model uses a binary, right/wrong categorization of the items (rather than the exact answer) to predict the long-run outcome of interest and compares that version of an item-based model to the predictive power of the average. Panel (b) of appendix figure A.4 is also identical to figure 5 from the main text, except that the benchmark model uses the scaled scores (theta) from the Texas IRT model (rather than the average score) to predict the long-run outcome of interest. In both cases, we find results that are substantively and quantitatively similar to the results in the main text: Models that use items (binary or the exact answer) outperform the benchmark (average test scores or IRT scaled scores).

[Figure A.4 about here.]

A.10 Split sample IV tests

Students. The single latent factor model implies that predicted values formed using items should be identical to predicted values formed using aggregates (i.e. that they should land on the 45 degree line if plotted against one another). In section 6.1, we show visual evidence (figure 6) that the predicted values formed using items often disagree with the predicted values formed using aggregates. However, some of this disagreement is surely due to sampling variation.

Thus, in this appendix, we formally test a related claim: that the predicted values are not even similar to one another *on average*. Clearly, if the predicted values are different on average (in the sense that the regression slope in figure 6 does not equal 1), then it cannot be the case that with infinite data every individual observation would land on the 45 degree line; and hence, the single latent factor model cannot be true.

To formally test this claim, we split the data into two folds. We use fold one to train our ML algorithm and form predicted values within that fold ($\hat{Y}_{agg}, \hat{Y}_{item,1}$). We then independently train a second item-level model in fold two, and use that second model to predict student outcomes in the first fold ($\hat{Y}_{item,2}$). Because the models are trained independently, the measurement error will also

be independent across the item-level predicted values: $\hat{Y}_{item,1}$ and $\hat{Y}_{item,2}$. This allows us to account for the measurement error and correct the resulting attenuation bias by estimating the following 2SLS model: $\hat{Y}_{agg} = \beta_0 + \beta_1 \hat{Y}_{item,1} + \epsilon$, with first stage given by $\hat{Y}_{item,1} = \pi_0 + \pi_1 \hat{Y}_{item,2} + u$ and where $u \perp \epsilon$ by construction.

The results are contained in appendix table A.1. We can reject that the slope equals one for every outcome and at every grade level.

Teachers. Similar to the analysis with students, the single latent factor model implies that teacher value-added on predicted values formed using items should be identical to teacher value-added on predicted values formed using aggregates (i.e. that they should land on the 45 degree line if plotted against one another). In section 6.2, we show visual evidence (figure 7) that teacher value-added formed using items often disagrees with teacher value-add using aggregates. However, some of this disagreement is surely due to sampling variation.

Thus, in this appendix, we formally test a related claim: that the teacher value-added estimates are not even similar to one another *on average*. Clearly, if the estimates are different on average (in the sense that the regression slope in figure 7 does not equal 1), then it cannot be the case that, with infinite data, every individual teacher would land on the 45 degree line; and hence, the single latent factor model cannot be true.

To formally test this claim, we split the data into two folds *within teacher*. We next estimate teacher value-add within each fold to generate two, independent estimates of teacher value add: $\hat{\delta}_{j(i,g),item,1}$ and $\hat{\delta}_{j(i,g),item,2}$. Because the two estimates are generated on different sets of students and adjusted using the standard jack-knife, empirical-Bayes method to eliminate the common classroom shock, we know that the two estimates will have orthogonal measurement error. Thus we can account for the measurement error in the regression slope and correct the resulting attenuation bias by estimating the following 2SLS model: $\hat{\delta}_{agg} = \beta_0 + \beta_1 \hat{\delta}_{item,1} + \epsilon$, with first stage given by $\hat{\delta}_{item,1} = \pi_0 + \pi_1 \hat{\delta}_{item,2} + u$ and where $u \perp \epsilon$ by construction.

The results are contained in appendix table A.2. We can reject that the slope equals one for every outcome and at every grade level.

[Table A.1 about here.]

[Table A.2 about here.]

A.11 Robustness to using empirical quantiles

To highlight disagreement at the bottom of the distribution between predicted values formed using items versus aggregates (and teacher value-add on those same predicted values), our preferred approach in sections 6.1 and section 6.2 follows Petek and Pope (2023) and Rose et al. (2022) by standardizing the predictions (or, in the case of teachers, TVA on the predicted values) to have mean 0 and standard deviation 1, and then counting a student (or teacher) in the bottom 5% of the relevant distribution if the value falls below -1.645. This approach will be exactly correct under the assumption that the estimates are normally distributed, and has the advantage that it simplifies inference considerably relative to defining the bottom 5% using the corresponding empirical quantile. However, out of an abundance of caution, in Appendix Tables A.3 and A.4 we redo this exercise defining the bottom 5 percent using the empirical 5th quantile instead. The results are

virtually unchanged.

[Table A.3 about here.]

[Table A.4 about here.]

A.12 Details of bootstrap procedure

Bootstrap algorithm for disagreement shares in table 4

Formally, the bootstrap algorithm proceeds as follows:

1. Randomly split the data into train (75%) and test (25%) folds.
2. Estimate models (10) and (11) in the training fold.
3. Use the estimated models to form predicted values using aggregate scores (\hat{Y}_i) and items (\tilde{Y}_i) in the hold-out sample.
4. Estimate the standard error of the predicted values by regressing the outcome on the predicted values out of sample in the test fold. Call these standard errors $SE_{\hat{Y}_i}$ and $SE_{\tilde{Y}_i}$.
5. Draw “placebo” predicted values for items and aggregate scores under the null that the predicted value for the item based model is identical to the predicted value from the aggregate based model. Formally, for the aggregate score model we draw the placebo predicted value as: $x_i \sim N(\hat{Y}_i, SE_{\hat{Y}_i})$. For the item based model we draw the placebo predicted value as: $z_i \sim N(\tilde{Y}_i, SE_{\tilde{Y}_i})$.
6. Calculate the bottom 5% disagreement share using the placebo predicted values x_i and z_i which, by construction, can only differ from one another as a result of the measurement error in the predicted value of the respective random forest models.
7. Repeat steps 5 and 6 for 1,000 iterations.
8. Calculate the bootstrap p-value as the share of placebo disagreement shares that are larger than the disagreement share observed in the actual data (specifically, the fraction of placebo disagreement shares larger than the corresponding number in the 12-17 column of table 4).

Bootstrap algorithm for disagreement shares in table 5

Formally, the bootstrap algorithm proceeds in two stages. In stage one, we estimate a bootstrap standard error for each teacher’s value-add that incorporates the sampling variability from the random forest estimation step.⁵¹ In stage 2, we generate a distribution of disagreement shares under the null of no-disagreement using the sampling distribution implied by stage 1.

Stage 1: Estimating the sampling variability in teacher value-add

1. Use the predicted values and standard errors from the “student” bootstrap algorithm described earlier in this appendix to draw “perturbed” predicted values: For the aggregate score based model, we draw these placebo values as $x_i \sim N(\hat{Y}_i, SE_{\hat{Y}_i})$. For the item based model we draw the placebo predicted value as: $z_i \sim N(\tilde{Y}_i, SE_{\tilde{Y}_i})$.
2. Estimate jack-knife, empirical-bayes estimates of teacher value-add using the perturbed val-

⁵¹Note that this is not trivial, since we are using the jack-knife empirical-bayes procedure from Chetty et al. (2014a).

ues as the outcome: $(\hat{\delta}_j, \tilde{\delta}_j)$

3. Repeat steps 1 and 2 for 1,000 iterations.
4. Calculate the standard deviation of the distribution of item level estimates for each teacher's value-add (SD_j^{item}) and the standard deviation of the distribution of aggregate score based estimates for each teacher's value-add (SD_j^{agg}). These parameters quantify the sampling variability in the teacher value-add estimates emerging from the random forest estimation step.

Stage 2: Estimating a bootstrap p-value for the disagreement share

1. Draw placebo estimate of teacher value-add under the null that the teacher's value add using predicted values from items is equal to their value-add found using predicted values from the aggregate score: $x_j \sim N(\hat{\delta}_j, SD_j^{agg})$ for the estimate based on the aggregate score, and $z_j \sim N(\hat{\delta}_j, SD_j^{item})$ for the estimate based on the items.
2. Calculate the bottom 5% disagreement share using the placebo teacher value-add estimates x_i and z_i which, by construction, can only differ from one another as a result noise.
3. Repeat steps 1 and 2 for 1,000 iterations.
4. Calculate the bootstrap p-value as the share of placebo disagreement shares that are larger than the disagreement share observed in the actual data (specifically, the fraction of placebo disagreement shares larger than the corresponding number in the 12-17 column of table 5).

A.13 Additional tables and figures

[Table A.5 about here.]

[Figure A.5 about here.]

[Table A.6 about here.]

[Table A.7 about here.]

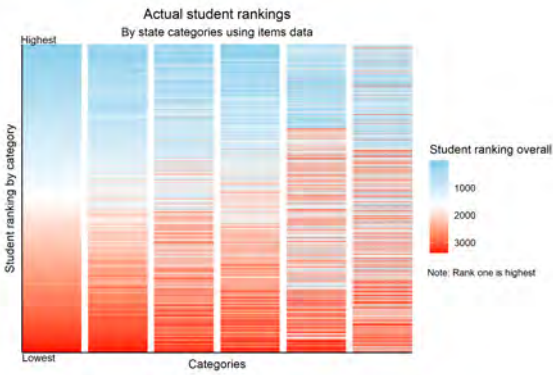
[Figure A.6 about here.]

B Proof of Proposition 1

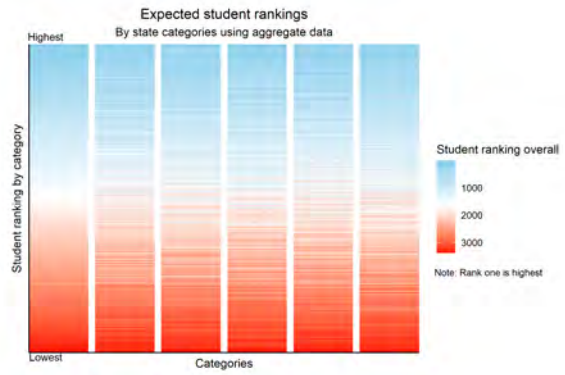
To show that $1 \Rightarrow 2$, first assume that the single latent factor model is true. Note that the single latent factor, θ_i , is identified as long as $|T| > 2$ (Cunha et al., 2010). Consider the corresponding aggregator: $a(D_i) = \theta_i$. Observe that the single latent factor model implies that, for any decision \mathbb{D} and associated prediction exercise (Y), we must have that $D_i | \theta_i \perp Y_i$. Therefore, it must be that for any alternative aggregator such that $\tilde{a}(D_i) \neq \theta_i$, we have that $\text{MSE}_Y(\theta) \leq \text{MSE}_Y(\tilde{a})$, which shows that $a(D_i) = \theta_i$ is optimal for every decision.

To show that $2 \Rightarrow 1$, we consider two decisions with observable features Z and Y and assume by way of contradiction that there is more than one non-trivial latent factor (θ_1 and θ_2), where by non-trivial we have in mind that the latent factors are differently predictive with respect to Y and Z : $\text{MSE}_Y(\mathbb{E}(Z|\theta_1, \theta_2)) > \text{MSE}_Y(\mathbb{E}(Y|\theta_1, \theta_2))$. Observe that $D_i | (\theta_{1i}, \theta_{2i}) \perp Y_i, Z_i$. Thus, the optimal aggregator for Y must be such that $a^*(D_i) = \mathbb{E}(Y|D_i) = \mathbb{E}(Y|\theta_{1i}, \theta_{2i})$. Now suppose there is a single optimal aggregator. Then it must also be that $a^*(D_i) = \mathbb{E}(Z|D_i) = \mathbb{E}(Z|\theta_{1i}, \theta_{2i})$ and hence we have shown that, if a single optimal aggregator exists, $\mathbb{E}(Y|\theta_{1i}, \theta_{2i}) = \mathbb{E}(Z|\theta_{1i}, \theta_{2i})$. But this contradicts our assumption that there are two “non-trivial” latent factors. Hence, the single latent factor model must hold.

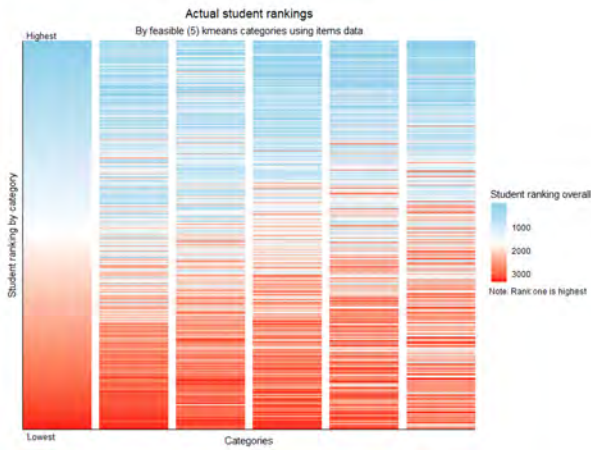
(a) State Categories: Actual



(b) State Categories: Placebo



(c) Categories from Section 7: Actual



(d) Categories from Section 7: Placebo

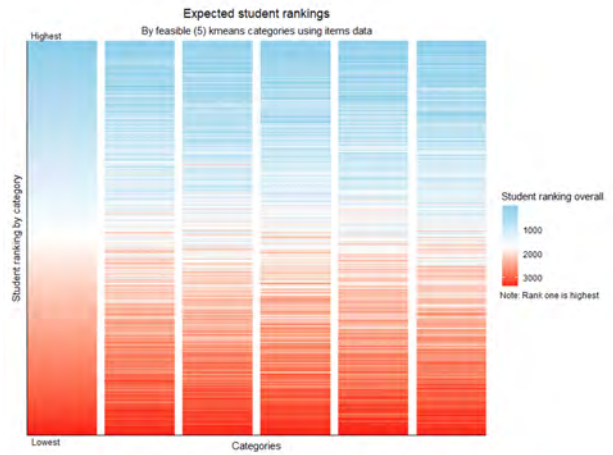
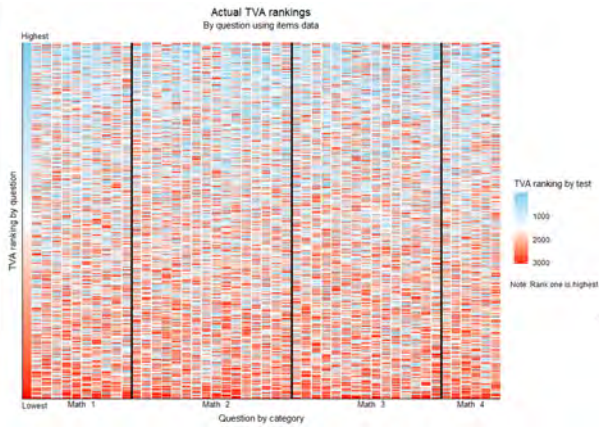


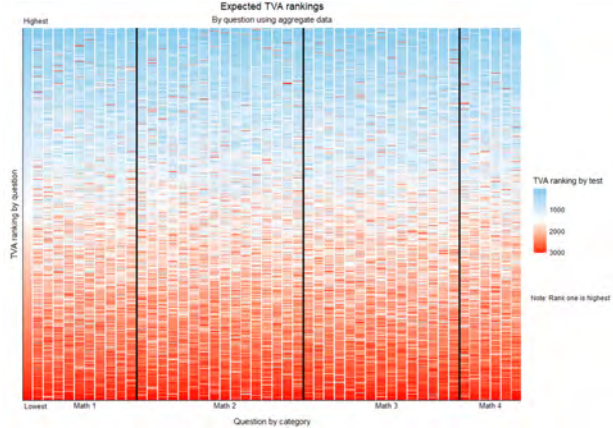
Figure A.1: Hidden structure in student performance: Robustness to alternative categorizations

Notes: These figures are identical to those contained in figure 3 from the main text, except that they use different categorizations of the underlying items. More precisely: the categories used to construct panels (a) and (b) of this figure are created using off-the-shelf content categories provided by the state of Texas; while the categories used to construct panels (c) and (d) are constructed using the item-linking algorithm developed in section 7. In both cases, we find substantively similar results.

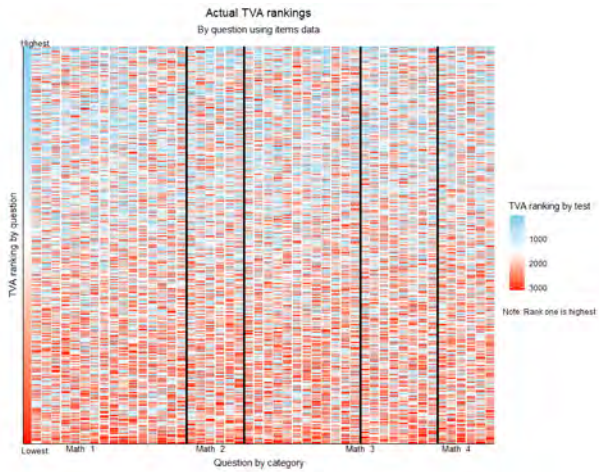
(a) Actual distribution: no shock adjustment



(b) Placebo distribution: no shock adjustment



(c) Actual distribution: lag-score interactions



(d) Placebo distribution: lag-score interactions

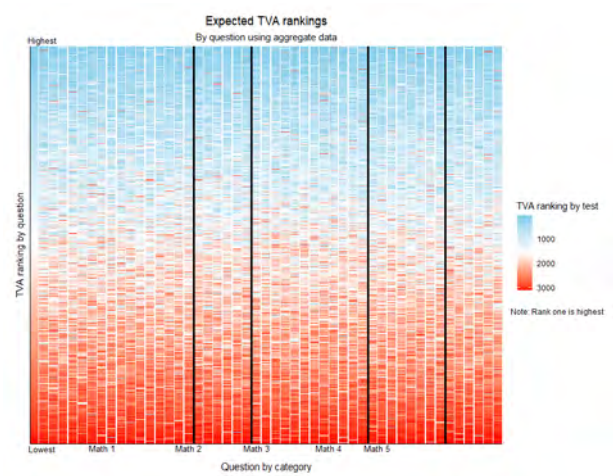


Figure A.2: Robustness: TVA rank-swapping

Notes: These figures are identical to figure 4 from the main text except that in panels (a) and (b) the fixed effects from model 8 are not adjusted for the classroom shock; and, in panels (c) and (d), the value-added measures are generated using models that include baseline-score by question interactions as would be implied by equation 7.

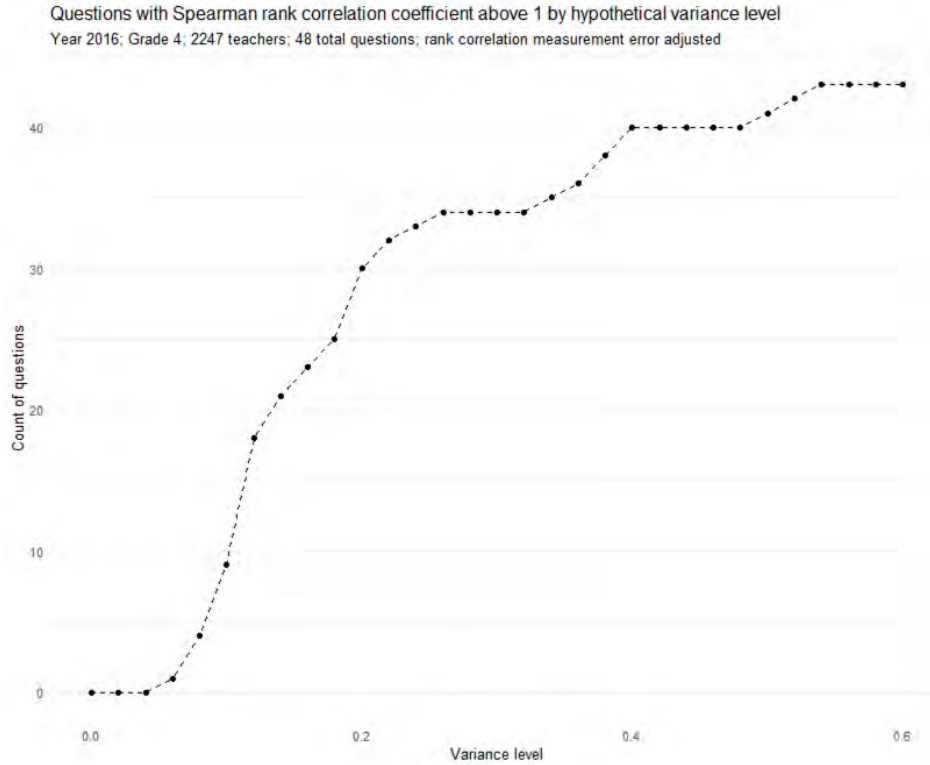


Figure A.3: Sensitivity to measurement error: Rank correlations

Notes: This figure displays the number of item level teacher-value add estimates exhibiting a rank-correlation with traditional value-add that is different from one under different assumptions about the degree of measurement error (Kitagawa et al., 2018). Specifically, we plot the assumed proportion of observed variance attributable to measurement error (x-axis) against the number of items where we would fail to reject that the Spearman rank correlation equals one using the Kitagawa et al. (2018) bias-correction method. We find that there are items exhibiting a statistically significant deviation from one even if we assume that over 50% of the variation in teacher value-added is due to measurement error. This suggests our conclusions are robust to large degrees of measurement error.

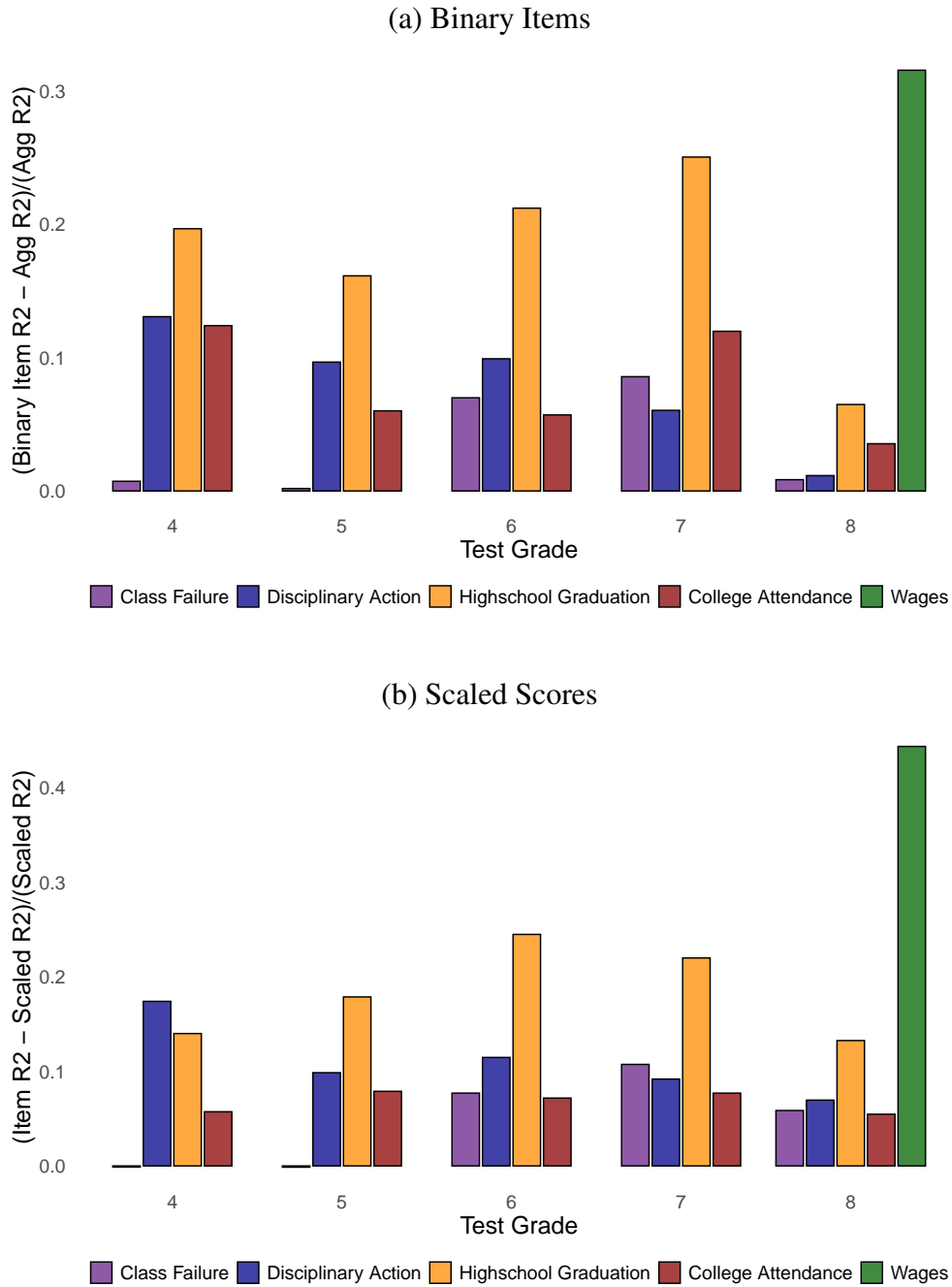


Figure A.4: Robustness: Hidden information about decisions

Notes: This figure shows that the results contained in figure 5 are robust to both: (1) alternative ways of using the items to predict aggregate achievement; and (2) alternative ways of representing aggregate achievement. Panel (a) is identical to figure 5 from the main text, except that the underlying random forest model uses a binary, right/wrong categorization of the items (rather than the exact answer) to predict the long-run outcome of interest. Panel (b) is also identical to figure 5 from the main text, except that the underlying random forest model uses the scaled scores (theta) from the Texas IRT model (rather than the average score) to predict the long-run outcome of interest. In both cases, we find results that are substantively and quantitatively similar.

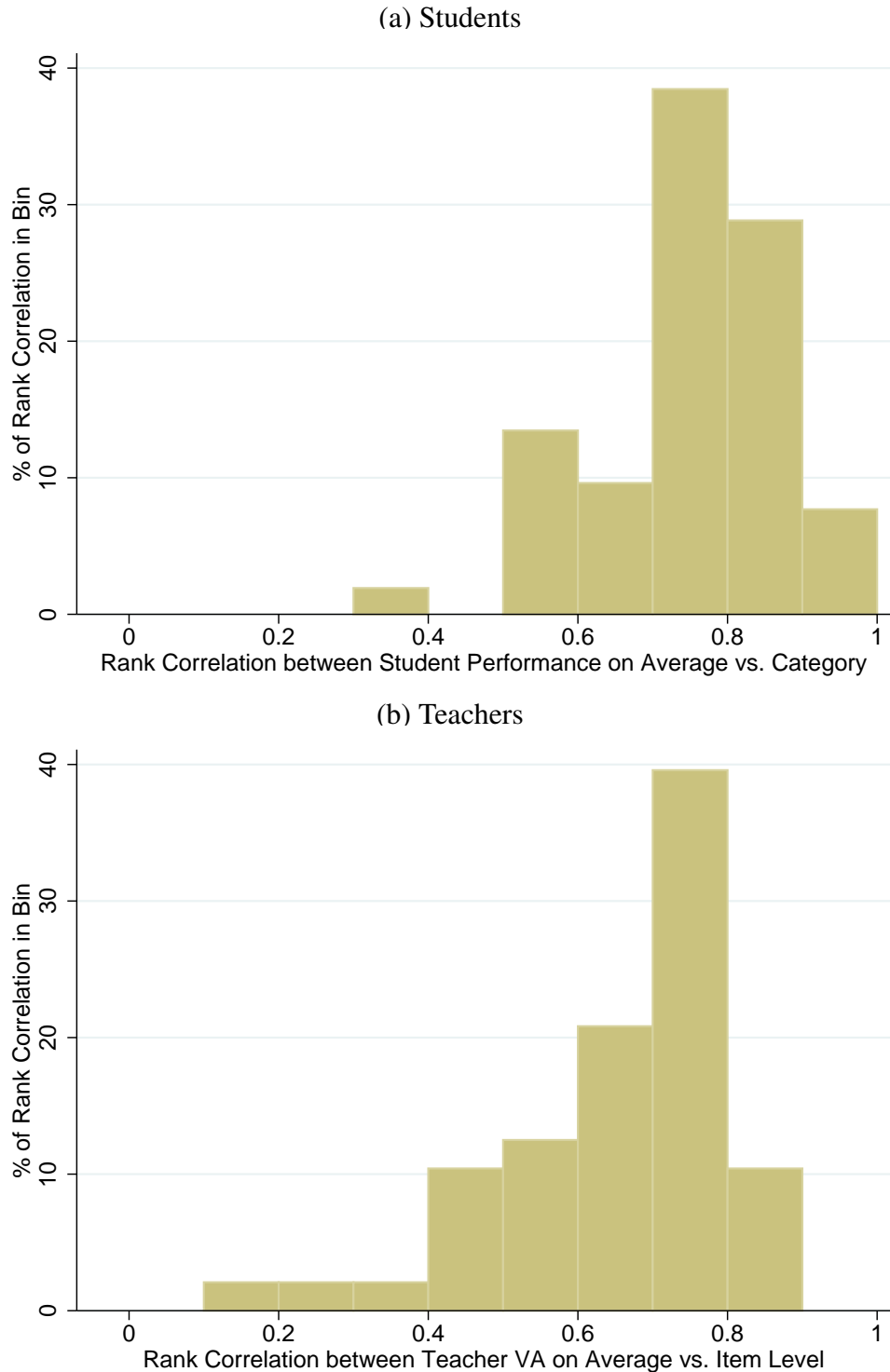


Figure A.5: Distribution of Rank Correlations between Items and Aggregates

Notes: These figures show the distribution of rank correlations when using aggregates versus items. Figure A.5(a) displays the distribution of rank correlations of a student’s average performance on a test and the 5 test sub-categories (categories are constructed using k -means clustering). This figure includes grade 4-8 tests in 2017 and 2018 and so groups 50 rank correlations. The average rank correlation we find is 0.75. Similarly, Figure A.5(b) shows the distribution of rank correlations for a teacher’s VA on the average test score and a teacher’s VA on each individual item. We do this exercise for one test (grade 4 in 2016) and so the histogram groups 50 rank correlations. As in Figure 4, the ranks are calculated among 2,280 fourth grade teachers. The average rank correlation we find is 0.66.

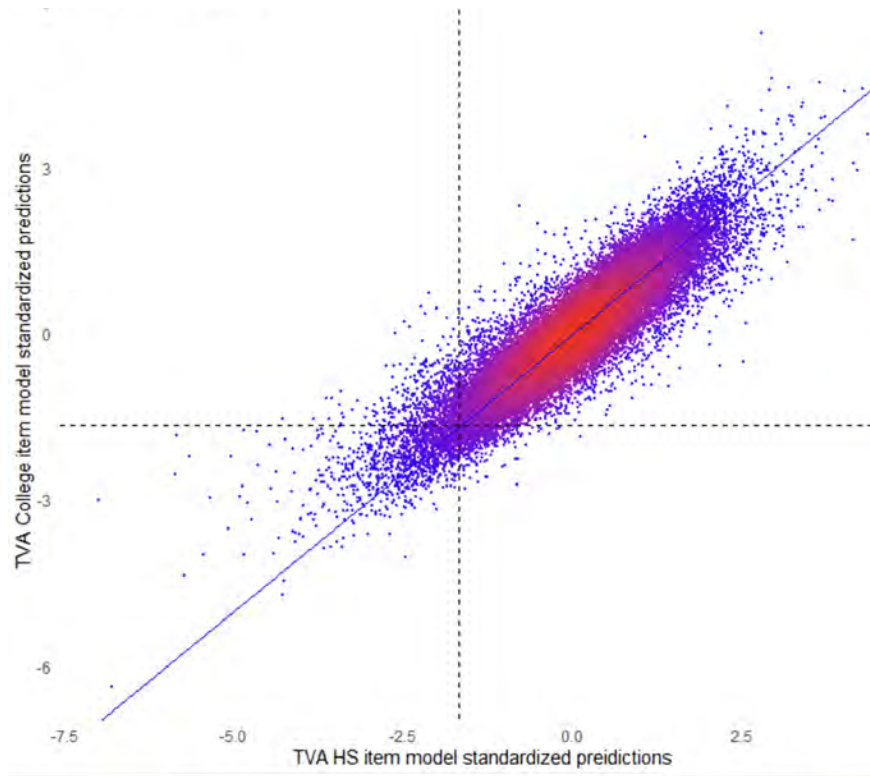


Figure A.6: Items vary in the amount of signal they hold for teacher value-add on different long-run outcomes

Notes: This figure demonstrates that different test items carry different amounts of signal about teacher value-add on different outcomes. In this figure, the x-axis denotes teacher value-add on item-level predicted values for high-school graduation as the left hand side variable. The y-axis denotes teacher value-add on item-level predicted values from model (10) with college attendance as the left hand side variable. Thus each observation is a teacher. Predicted values are standardized within year-and-outcome to have mean zero, standard deviation 1 prior to plotting. The wide dispersion in the data implies that some items carry different predictive power over different outcomes and hence there cannot be a single, optimal aggregator.

Table A.1: 2SLS Predicted values from item model on the aggregates (students)

	2012	2013	2014	2015	2016	2017	2018	2019	12-17	P-Value	Obs.
<i>Panel A: Predicted graduation</i>											
Grade 4	0.86	0.85							0.86	0	114,749
Grade 5	0.91	0.86	0.87						0.88	0	177,437
Grade 6	0.82	0.80	0.87	0.85					0.84	0	231,293
Grade 7	0.78	0.78	0.83	0.80	0.84				0.82	0	283,090
Grade 8	0.82	0.87	0.84	0.90	0.91	0.94			0.92	0	344,113
<i>Panel B: Predicted college</i>											
Grade 4	0.91								0.91	0	57,990
Grade 5	0.95	0.93							0.94	0	118,376
Grade 6	0.95	0.93	0.94						0.94	0	172,817
Grade 7	0.88	0.87	0.91	0.89					0.89	0	223,293
Grade 8	0.88	0.93	0.91	0.96	0.94				0.93	0	276,778
<i>Panel C: Predicted wages</i>											
Grade 8	0.92								0.92	0	39,284
<i>Panel D: Predicted disc. infr.</i>											
Test grade 4	0.83	0.80	0.83	0.84	0.88	0.85	0.83		0.84	0	403,569
Test grade 5	0.90	0.89	0.87	0.92	0.91	0.93	0.91		0.91	0	409,260
Test grade 6	0.91	0.92	0.91	0.89	0.90	0.89	0.92		0.91	0	404,919
Test grade 7	0.86	0.87	0.87	0.85	0.86	0.87	0.91		0.87	0	396,855
Test grade 8	0.85	0.94	0.89	0.95	0.94	0.91	0.94		0.92	0	406,013
<i>Panel E: Predicted pass/fail class</i>											
Grade 4	0.92	0.94	0.94	0.92	0.95	0.93	0.94	0.96	0.94	0	2,078,976
Grade 5	0.96	0.94	0.95	0.97	1.01	0.97	1.01	0.99	0.97	0	2,159,728
Grade 6	0.88	0.94	0.94	0.91	0.90	0.90	0.91	0.94	0.92	0	1,251,104
Grade 7	0.91	0.92	0.94	0.92	0.87	0.89	0.92	0.91	0.91	0	1,967,294
Grade 8	0.87	0.97	0.92	0.94	0.94	0.95	0.96	0.95	0.94	0	1,723,489

Notes: This table provides evidence that the points in figure 6 do not lie on the 45 degree line. To formally test this claim, we split the data into two folds. We use fold one to train our ML algorithm and form predicted values within that fold ($\hat{Y}_{agg}, \hat{Y}_{item,1}$). We then independently train a second item-level model in fold two, and use that second model to predict student outcomes in the first fold ($\hat{Y}_{item,2}$). Because the models are trained independently, the measurement error will also be independent across the item-level predicted values: $\hat{Y}_{item,1}$ and $\hat{Y}_{item,2}$. This allows us to account for the measurement error and correct the resulting attenuation bias by estimating the following 2SLS model: $\hat{Y}_{agg} = \beta_0 + \beta_1 \hat{Y}_{item,1} + \epsilon$, with first stage given by $\hat{Y}_{item,1} = \pi_0 + \pi_1 \hat{Y}_{item,2} + u$ and where $u \perp \epsilon$ by construction. This table presents slope coefficients from this regression for every year-outcome-grade level combination. The column labeled “12-17” pools across years, within an outcome-grade level. The p-value is from a formal test of whether the coefficient in the “12-17” column equals 1. We reject that the slope equals one for every outcome and at every grade level.

Table A.2: 2SLS Predicted TVA values from item model on the aggregates (teachers)

	2013	2014	2015	2016	2017	2018	13-18	P-Value	Obs.
<i>Panel A: Predicted graduation</i>									
Grade 5	1.06						1.06	0	1,845
Grade 6	1.01	1.12					1.11	0	2,424
Grade 7	0.92	1.15	1.20				1.02	0	3,376
Grade 8	1.05	1.10	1.58	1.29			1.08	0	5,010
<i>Panel B: Predicted college</i>									
Grade 6	1.12						1.12	0	975
Grade 7	0.93	1.10					0.89	0	1,890
Grade 8	1.06	1.21	1.45				1.23	0	3,361
<i>Panel C: Predicted disc. infr.</i>									
Test grade 4	1.39	1.39	1.55	1.33	1.32		1.12	0	11,861
Test grade 5	1.21	1.13	1.14	1.18	1.23		1.09	0	9,898
Test grade 6	1.09	1.21	1.07	0.94	0.96		1.03	0	7,185
Test grade 7	0.98	1.07	1.17	1.30	1.19		0.94	0	6,348
Test grade 8	1.05	1.04	1.31	1.36	1.15		1.19	0	6,596
<i>Panel D: Predicted pass/fail class</i>									
Grade 4	1.27	1.32	1.14	1.19	1.11	1.19	1.01	0	12,775
Grade 5	1.18	1.24	1.26	1.27	1.16	1.26	1.04	0	10,947
Grade 6	0.94	1.01	0.97	0.90	0.96	0.98	1.07	0	8,706
Grade 7	0.83	1.07	0.96	0.83	0.77	0.83	1.08	0	7,940
Grade 8	0.91	0.86	1.07	1.13	1.06	1.16	1.12	0	8,140

Notes: This table provides evidence that the points in figure 7 do not lie on the 45 degree line. To formally test this claim, we split the data into two folds *within teacher*. We next estimate teacher value-add within each fold to generate two, independent estimates of teacher value add: $\hat{\delta}_{j(i,g),item,1}$ and $\hat{\delta}_{j(i,g),item,2}$. Because the two estimates are generated on different sets of students and adjusted using the standard jack-knife, empirical-Bayes method to eliminate the common classroom shock, we know that the two estimates will have orthogonal measurement error. Thus we can account for the measurement error in the regression slope and correct the resulting attenuation bias by estimating the following 2SLS model: $\hat{\delta}_{agg} = \beta_0 + \beta_1 \hat{\delta}_{item,1} + \epsilon$, with first stage given by $\hat{\delta}_{item,1} = \pi_0 + \pi_1 \hat{\delta}_{item,2} + u$ and where $u \perp \epsilon$ by construction. The column labeled “13-18” pools across years, within an outcome-grade level. The p-value is from a formal test of whether the coefficient in the “13-18” column equals 1. We reject that the slope equals one for every outcome and at every grade level.

Table A.3: Robustness to Defining the Bottom 5% Using Empirical 5th Quantile (Students)

	2012	2013	2014	2015	2016	2017	2018	2019	12-19	Obs.
<i>Panel A: Predicted graduation</i>										
Grade 4	54.2%	49.3%							48.4%	114,761
Grade 5	47.8%	52.4%	51.7%						50.4%	177,415
Grade 6	55.9%	58.5%	55.6%	54.0%					55.8%	230,523
Grade 7	55.2%	55.5%	53.1%	53.3%	58.4%				55.1%	279,675
Grade 8	52.6%	51.6%	54.0%	46.6%	45.7%	50.7%			50.1%	320,374
<i>Panel B: Predicted college</i>										
Grade 4	45.9%								45.9%	57,978
Grade 5	47.7%	50.3%							45.9%	118,360
Grade 6	51.0%	55.1%	50.8%						49.0%	172,082
Grade 7	52.0%	52.4%	49.7%	47.9%					51.3%	221,763
Grade 8	49.8%	46.4%	50.7%	46.0%	43.8%				47.1%	258,497
<i>Panel C: Predicted wages</i>										
Grade 8	62.1%								62.1%	39,284
<i>Panel D: Predicted disciplinary infraction</i>										
Test grade 4	50.5%	50.2%	47.6%	44.3%	48.5%	55.9%	55.9%		50.4%	403,639
Test grade 5	38.1%	48.4%	43.9%	43.7%	43.0%	40.8%	40.2%		44.8%	409,265
Test grade 6	41.1%	47.4%	45.9%	43.0%	46.5%	45.6%	47.8%		49.1%	403,907
Test grade 7	57.3%	55.3%	57.9%	57.7%	58.7%	69.6%	58.5%		64.4%	391,744
Test grade 8	48.5%	49.5%	48.1%	79.0%	57.7%	57.3%	58.1%		61.2%	376,267
<i>Panel E: Predicted pass class</i>										
Grade 4	36.8%	33.3%	34.6%	35.8%	34.1%	34.0%	34.2%	36.5%	35.0%	2,078,976
Grade 5	39.4%	37.8%	35.0%	36.5%	34.3%	36.4%	34.6%	32.1%	35.8%	2,159,728
Grade 6	50.8%	48.9%	46.0%	51.4%	53.8%	53.9%	52.5%	51.6%	51.1%	1,251,104
Grade 7	51.6%	51.6%	48.0%	52.9%	54.6%	56.1%	50.5%	53.9%	52.6%	1,967,294
Grade 8	55.0%	51.6%	54.9%	49.2%	52.8%	52.8%	52.7%	53.3%	53.6%	1,723,489

Notes: This table is identical to table 4 in the main text, except that we define the bottom 5% using the empirical 5th quantiles. The results are virtually unchanged.

Table A.4: Robustness to Defining the Bottom 5% Using Empirical 5th Quantile (Teachers)

	2012	2013	2014	2015	2016	2017	2018	2019	12-19	Obs
<i>Teacher value-added on:</i>										
Predicted graduation	44.3%	42.1%	48.6%	41.3%	43.8%	40.0%			44.9%	68,933
Predicted college attendance	41.6%	39.1%	40.6%	41.6%	34.6%				39.6%	42,007
Predicted disciplinary infraction	41.3%	42.0%	41.7%	39.6%	42.6%	42.8%	44.7%		42.1%	148,105
Predicted pass/fail class	38.3%	40.7%	40.4%	38.8%	37.8%	38.9%	37.1%	36.8%	38.5%	608,000

Notes: This table is identical to table 5 in the main text, except that we define the bottom 5% using the empirical 5th quantiles. The results are virtually unchanged.

Table A.5: Coverage of Long-Term Data Linkage

	Last Year of Data (1)	Cohorts Covered ^a (2)	Match Rate for Test-Taker Sample (3)	Match Rate for VA Sample (4)
Graduation ^b	2011-12 to 2012-13	Entering 4 th grade in 2012-13 or before	55% (332,643 of 607,826)	
College-Going ^b	2011-12 to 2011-12	Entering 4 th grade in 2012-13 or before	55% (332,643 of 607,826)	
High School Outcomes: ^c				
Absences	2003-04 to 2016-17	N/A	Entering 3 rd grade in 2008-09 or before	72% (639,035 of 885,235)
GPA	2003-04 to 2016-17	N/A	Entering 3 rd grade in 2008-09 or before	69% (612,478 of 885,235)
'Effort' GPA	2003-04 to 2016-17	N/A	Entering 3 rd grade in 2008-09 or before	56% (490,983 of 885,235)
Days Suspended	2003-04 to 2016-17	N/A	Entering 3 rd grade in 2008-09 or before	74% (653,005 of 885,235)
Held Back	2003-04 to 2016-17	N/A	Entering 3 rd grade in 2008-09 or before	68% (601,338 of 885,235)

^a For example, if third grade cohorts from 2007-08 and before are covered, then the linkage will include: third grade students 2003-04 through 2007-08, fourth grade students 2003-04 through 2008-09, and fifth grade students 2003-04 through 2009-10.

^b Graduation is coded as one if a student graduates or receives a special education certificate and zero if the student is coded as a dropout in the graduation data files. If the student is not present in these data files the student is coded as missing. Graduation also omits the 2003-04 fifth grade cohort as this cohort is not covered since the data does not start until 2011-12.

^c For high school outcomes we require the student's cohort to reach the end of eleventh grade by 2016-17. To be in these data we must observe a high school transcript (grades 9-12) for the student for at least one term. Days suspended are total days suspended during a student's high school career, with these data being universal (conditional on observing at least one transcript) since we assume you were not suspended if you have a transcript but no suspension record. Absences similarly record the total number of absences over a student's high school career, although we note that there are some students who have a transcript but are missing absence data. GPA is the average GPA over a student's high school career, although we cannot construct GPA for some student-term observations due to letter grades not being recorded in the transcript (e.g., grades coded as pass/fail); 'effort' GPA is constructed analogously. Held back is constructed as an indicator variable when a student's high school grade is the same in two consecutive years; this requires us to observe valid transcripts for two consecutive years and so data coverage is not universal for this outcome.

Table A.6: Classification accuracy: items and aggregates

		$Y_{agg} = 1$	$Y_{agg} = 0$
$Y_{item} = 1$	actual	0.831	0.544
	item	0.830	0.573
	agg	0.828	0.471
$Y_{item} = 0$	actual	0.477	0.442
	item	0.438	0.406
	agg	0.578	0.459

Notes: This table plots actual graduation rates, expected graduations using the item-based model, and expected graduation rates using the aggregate score-based model, for cases where the item-based model and the aggregate score-based model agree and disagree. Specifically, $Y_{agg} = 1$ denotes cases where the aggregate score based model predicts a greater than 50% likelihood of the student graduating, with $Y_{agg} = 0$ denoting the opposite. Y_{item} is defined identically. Rows labeled “actual” give the actual graduation rates of students in that cell. Rows labeled “item” give average graduation rates using the classification from the item based model. Rows labeled “agg” give average graduation rates using the classification from the aggregate based model. While both models have similar accuracy in cases where they agree, the item model gives expected graduation rates that are closer to actual graduation rates in cases where they disagree.

Table A.7: Formal Tests of “No Information Gain” Null Hypothesis

	$p = P(F \geq f \mid H_0 : \text{MSE}(Y, \hat{Y}_{\text{item}}) \geq \text{MSE}(Y, \hat{Y}_{\text{avg}}))$					
	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	All grades (Joint Test)
Wages					0.39	0.39
College	0.04	0.06	0.03	0.00	0.00	0.00
High School Graduation	0.25	0.12	0.02	0.00	0.00	0.00
Discipline Infraction	0.06	0.03	0.00	0.01	0.02	0.00
Course Failure	0.43	0.66	0.03	0.04	0.16	0.01
<i>Omnibus test (all grades and all outcomes):</i>						0.00

Notes: This table shows that the extra information in the raw test items is connected to important long-run outcomes. Formally, this table displays p-values from tests of the “no information gain” null hypothesis described in section 6.1. Jointly tested across all grades and long-run outcomes, we can reject the null of no information gain. Moreover, we can also reject the null of no information gain for nearly all (22 out of 26) of the individual grade-by-outcome combinations.