Accountability, test prep incentives, and the design of math and English exams

Evan Riehl*
Meredith Welch+

March 2022

Abstract. We examine how incentives for test prep varied between math and English language arts (ELA) on U.S. state accountability exams. We collected data on exam structure for grade 3–8 tests in six states that are the setting for most U.S. research in literatures where accountability matters. We show that math exams typically measured ability more precisely for students on the margin of achieving proficiency. This gave educators an incentive to spend more time preparing students for math tests than for ELA tests, consistent with the common finding of larger math effects in these literatures.

*Riehl (corresponding author): Department of Economics, Cornell University, 266 Ives Hall, Ithaca, NY 14853 (email: eriehl@cornell.edu). +Welch: Department of Policy Analysis and Management, Cornell University, Martha Van Rensselaer Hall, Ithaca, NY 14853 (email: msw274@cornell.edu). We thank Peter Bergman, Julie Cullen, Joshua Goodman, Mike Lovenheim, Seth Sanders, and several anonymous referees for helpful comments. We are grateful to Morgan Leung and Sarah Mather for excellent assistance with the data collection and analysis. All errors are our own.

1. Introduction

Research in the economics of education often finds that accountability policies have a larger effect on math test scores than on English language arts (ELA) scores. For example, Dee and Jacob (2011) show that No Child Left Behind increased 4$^{\text{th}}$ grade math scores by 0.23 standard deviations, while ELA scores increased by only 0.06 SDs. Neal and Schanzenbach (2010), Rockoff and Turner (2010), and Rouse et al. (2013) each find that state accountability policies induced larger gains in math than in ELA.

This pattern also arises in other literatures where standardized testing and accountability are important. Abdulkadiroğlu et al. (2011) show that admission to Boston charter schools raised students' achievement by 0.42 SDs in math, and 0.25 SDs in ELA. Chetty et al. (2014a) find that the standard deviation of teacher value added in a large urban school district is roughly 50 percent larger in math than in ELA.

A common hypothesis for this pattern is that most math learning takes place at school, while students primarily learn English and reading at home. But Jackson et al. (2014) acknowledge: "There is no clear explanation for this fact."

This paper shows that this pattern can partly be explained by differences between math and ELA in the design of state accountability exams, which created stronger incentives for test prep in math. We collected data from the technical reports of grade 3–8 standardized exams in six states that are the setting for most research on education accountability in the United States. We show that, relative to ELA exams, math exams usually measured ability more precisely for students on the margin of achieving proficiency. This gave educators an incentive to spend more time preparing students for math tests than for ELA tests, consistent with the pattern of estimates in literatures where accountability matters.

We begin the paper with a brief review of math and ELA estimates in three research areas: 1) accountability policies (Figlio and Loeb, 2011); 2) admission to charter schools (Epple et al., 2016); and 3) variation in teacher value added (Koedel et al., 2015). Accountability through standardized testing is important in each of these literatures, as it impacts funding, the renewal of school charters, and teacher employment. We show that in each research area, there is a remarkably consistent pattern of larger effects in math than in ELA.

Next, we develop a framework that shows how a teacher's incentives for test prep depend on the structure of the accountability exam. Test prep increases students' expected exam performance, but requires costly effort. Our key assumptions are that teachers seek to maximize the proficiency rate in their class (a central metric in many accountability systems), and that they target instruction to marginally-proficient students (Neal and Schanzenbach, 2010). Under these assumptions, teachers exert more effort on test prep when there are

more students on the margin of proficiency, and when the exam is a more precise measure of ability for these marginal students.

We show that the design of the accountability exam affects these test prep incentives in two ways. First, test designers choose a proficiency standard, which affects how many students are near the proficiency margin. Exams with a "low bar" for achieving proficiency will tend to have fewer marginal students. Second, the exam's precision for the ability of marginally-proficient students depends on the types of questions. Exams measure ability more precisely when there are more questions at an appropriate difficulty level, and when questions are more "discriminating" between higher- and lower-ability test takers.

Our empirical analysis shows how state accountability exams varied in these dimensions during 2000–2008. Our data come from the technical reports of grade 3–8 math and ELA exams in Florida, Illinois, Massachusetts, New York, North Carolina, and Texas, which are the settings for most of the papers in our literature review. Each of these states implemented an accountability policy in the 1990s (Dee and Jacob, 2011), and accountability expanded in 2006 with the adoption of No Child Left Behind. The technical reports contain information on the structure of the accountability exams (e.g., the number of questions and their difficulty) and the realized distribution of test scores. We simulate data for each test to estimate the number of students near the proficiency margin, and the effect of an increase in these students' ability on their likelihood of achieving proficiency.

Our main finding is that test prep incentives were usually stronger in math than in ELA. Several differences in exam structure explain this fact. Math exams typically contained more questions at an appropriate difficulty level for marginally-proficient students, which made them a more precise measure of ability in this region. Math exams often had more questions overall, and these questions were more discriminating on average. Lastly, math exams often had lower proficiency rates than ELA exams, which meant that there were more students near the proficiency margin in the average classroom.

We show that variation in test prep incentives can explain a significant portion (but not all) of the variation between math and ELA effects in the literature. The math estimates in our literature review are roughly 50 percent larger than the ELA estimates on average, and the mean difference in test prep incentives in our preferred metric is 22 percent. We link the math/ELA ratio of point estimates in each paper to the math/ELA ratio of test prep incentives in the exams that overlap with the authors' samples. The two variables have a strong positive correlation: a 10 percent increase in the ratio of test prep incentives is associated with an eight percent increase in the ratio of literature estimates. This suggests that test prep is one reason why math estimates tend to be larger in the literature, although the magnitudes of our findings suggest that other factors play a role in this pattern.

The patterns we identify do not hold in every exam, but we think that they reflect general differences between math and ELA testing that are salient to teachers. It is often said that math requires cumulative knowledge, and most math questions can be described briefly. This may make it easier for test designers to write exams that precisely measure ability on the proficiency margin, and likewise make it easier for educators to prepare students for exams. We present evidence that exams in other U.S. states have a similar structure to those in our main sample, suggesting that our findings reflect general differences between math and ELA testing.

Our paper provides one explanation for the common finding that math scores are more responsive to accountability pressures than ELA scores. There is limited research that tries to explain this pattern. Many researchers hypothesize that it arises because school inputs matter more for achievement in math than in ELA (Jacob, 2005; Jackson et al., 2014), and our results do not rule out this possibility. Our findings are more closely related to Kane and Staiger (2012)'s hypothesis that ELA exams are less sensitive to teacher effort; we provide direct evidence on this claim and link it to incentives for test prep.[1] Our literature review focuses on papers that use data from the 1990s and 2000s, but the pattern of larger effects in math also arises in more recent contexts where educators face accountability (Jackson, 2018; Cohodes et al., 2021). By contrast, we show that math estimates tend to be slightly *smaller* than ELA estimates in an older literature on class size (Glass and Smith, 1979; Krueger, 1999), in which many papers use data from years with low accountability.

We contribute to research on the incentive effects of accountability policies in education. This work has shown how accountability affects the allocation of school resources (Jacob and Levitt, 2003; Figlio and Winicki, 2005; Craig et al., 2013; Reback et al., 2014), teacher effort (Taylor and Tyler, 2012; Aucejo et al., 2020), and the distribution of test scores (Reback, 2008; Springer, 2008; Hemelt, 2011; Macartney et al., 2021). This work often finds that accountability pressures have a larger effect on scores in high-stakes exams than on low-stakes exams (Jacob, 2005; Figlio and Rouse, 2006; Imberman and Lovenheim, 2015; Bergbauer et al., 2018), suggesting that educators teach to the test (Holmstrom and Milgrom, 1991). There is also a large body of education research that finds that accountability pressures cause teachers to engage in test prep (Stecher and Mitchell, 1995; Koretz et al., 1996; Jones et al., 1999; Pedulla et al., 2003; Jennings and Bearak, 2014).[2]

Our paper adds to this literature by showing how the effects of accountability policies depend on the structure of the exams. For critics of standardized testing, our findings might

[1] Our paper is also related to work that shows that the precision of test scores matters for the effectiveness and evaluation of accountability policies (Kane et al., 2002; Kane and Staiger, 2002; Chay et al., 2005).

[2] Other research in education discusses the implication of test prep for the evaluation of accountability policies (e.g., Fuller et al., 2007; Ho, 2008).

suggest that the pattern of larger math effects in the literature does not reflect larger gains in generalizable math learning. This is consistent with Chetty et al. (2014b)'s finding that future earnings are *less* related to teacher value added in math than in ELA. But since our data only include scores on accountability exams, we do not take a stand on whether test-prep learning is valuable outside of the exam.

Our broader takeaway is that researchers should be cognizant of how accountability pressures and exam design can impact their findings (Jacob and Rothstein, 2016; Nielsen, 2019). We conclude the paper with guidance for gauging how test prep incentives vary across exams.

The paper proceeds as follows. Section 2 presents a brief review of math and ELA estimates in literatures where accountability matters. Section 3 develops our theoretical framework. Section 4 describes our data, methods, and main results on the test prep incentives in math and ELA exams. Section 5 examines the relationship between test prep incentives and literature estimates in our sample. Section 6 concludes.

2. Literature estimates

Table 1 summarizes math and ELA estimates in three literatures where accountability through standardized testing is important: A) accountability policies; B) admission to charter schools; and C) teacher value added. To be systematic, we use a review paper on each topic to define the set of studies for Table 1. We include only papers that present effects on grade 3–8 test scores in *both* math and ELA. We also exclude papers with identification strategies that are regarded as less credible by most researchers in the economics of education (e.g., propensity score matching or student fixed effects). Column (A) lists the studies that meet these criteria. Columns (B)–(E) show the location, exam years, and grades for the sample in each paper. Columns (F)–(G) show the math and ELA estimates from the authors' preferred specification.[3]

Panel A includes papers on the effects of accountability policies cited in the review by Figlio and Loeb (2011). This includes work on state- or district-level accountability systems that existed prior to No Child Left Behind (NCLB), as well as national studies of the effects of NCLB. In this panel, the estimates in columns (F)–(G) reflect the effects of an increase in the stakes of an exam—either from the introduction of an accountability system or from increased pressure due to a school's poor prior performance.

Panel B shows papers on the effects of charter school attendance from the literature review in Epple et al. (2016). The math and ELA effects all come from lottery-based identification strategies, which compare the test scores of admission lottery winners and losers. Accountability is important for charter schools because they face more stringent requirements to

[3] Most estimates are in effect size units (i.e., test scores are normalized to mean 0/SD 1), but the units differ in a few papers. The math and ELA effects are always from the same specification and in the same units.

operate than traditional public schools. For example, Massachusetts charter schools must renew their charter every five years, and the process includes a review of proficiency rates on state exams (Abdulkadiroğlu et al., 2011). Additionally, teacher pay at charter schools is often directly tied to student exam performance (Dobbie and Fryer Jr, 2011).

Panel C includes papers in the review by Koedel et al. (2015) that estimate variation in teacher value added. In this panel, the estimates in columns (F)–(G) represent the standard deviation of the teacher value added distribution in math and ELA. The link between these effects and accountability is weaker than in the other two literatures, but test prep is still likely to play a role. Researchers can typically only estimate teacher value added in districts where accountability systems exist to collect such data. Part of what distinguishes a good teacher in this metric may be their effectiveness at preparing students for exams. Consistent with this, Corcoran et al. (2011) find that teacher value added varies more on high-stakes exams than on low-stakes exams, and Macartney et al. (2018) present evidence that teacher effort responds to accountability incentives.

In each literature, there is a remarkably consistent pattern of larger effects on math scores than on ELA scores. Column (H) shows that math effects are more than 50 percent larger than ELA effects on average, and they are larger in magnitude in all but two of the papers in our review. This pattern is most pronounced in the accountability papers, for which the math effects are 75 percent larger than the ELA effects on average. The math/ELA ratio of estimates also tends to be higher in grades 6–8 than in grades 3–5 (1.66 vs. 1.44). The rest of our paper examines the role of test prep incentives in explaining this pattern.

3. Framework

This section presents a simple model of a teacher's incentives for test prep and shows how the structure of an accountability exam affects these incentives.

3.1. **Optimal test prep.**

We consider a teacher with a class of students indexed by i, and we let α_i denote the ability of each student. We interpret α_i as the level of student ability that would arise in the absence of accountability pressures. Thus α_i is determined by many factors, including the student's innate ability, their prior education, and the instruction choices that the teacher would make in the absence of accountability.

If accountability is important, the teacher can further increase student ability by exerting effort, e^*. This term can also represent the amount of extra class time devoted to test prep as a result of accountability pressures. We assume that teacher effort increases student ability through the following learning production function:

$$\theta_i = \alpha_i + g(e^*)h(\alpha^* - \alpha_i). \tag{1}$$

Effort is costly, which may reflect teacher morale or less time available for other material, and has diminishing returns through the function $g(\cdot)$. The teacher must also choose a target instruction level, α^*, that affects the amount of skill accumulation for each student. We assume $h(\cdot)$ is a decreasing function of the absolute value of the difference between the target level, α^*, and the student's ability, α_i, and it equals zero if $|\alpha^* - \alpha_i|$ is large. In other words, students learn more when their own ability is close to the instruction level. Students who find the test prep too easy or too hard may not benefit from this preparation.[4]

Together, student ability and any added skill from test prep determine the student's skill at the time they sit for the exam, θ_i.[5] The term θ_i reflects the student's potential to perform well on the exam, and it is important to note that this skill may or may not be useful outside of the exam (Holmstrom and Milgrom, 1991). The term "test prep" often has a negative connotation, but our paper does not take a stand on the value of test prep since we do not use data on outcomes other than exam scores. This caveat also applies to most of the papers in Table 1, which use test scores as the main outcome variable.[6]

The exam consists of multiple questions indexed by $q = 1, \ldots, Q$, and the student's post-prep skill affects their exam performance through the equation

$$p_q(\theta_i) \equiv Pr[u_{iq} = 1|\theta_i]. \tag{2}$$

u_{iq} is an indicator equal to one if student i correctly answers question q. The term $p_q(\theta_i)$ defines the probability of a correct answer to question q as a function of student skill, θ_i.

We consider test scores defined by whether or not a student meets a proficiency standard:

$$\tau(R_i) = \mathbb{1}\{R_i \geq \underline{\mathrm{R}}\}, \quad \text{where } R_i = \sum_{q=1}^{Q} u_{iq}. \tag{3}$$

The student's test score, $\tau(R_i)$, is a function of their raw score, R_i, which is the total number of correct answers.[7] There are many potential transformations of raw scores, including scale scores (as often reported by states) and standardized scores (as often used by researchers). We focus on an indicator for achieving *proficiency*, which occurs when the raw score, R_i,

[4] Our specification of student learning (equation 1) is similar to that in Duflo et al. (2011), except the authors also allow for the possibility of peer effects.

[5] Throughout the paper we use the terms "ability" and "skill" interchangeably to refer to both α_i and θ_i. The term θ_i might more accurately be called "test skill." But psychometricians often use "ability" or "skill" to refer to the object of measurement in writing exams, and so we follow their language.

[6] A notable exception is Chetty et al. (2014a), who examine the predictive power of teacher value added for long-run earnings in a companion paper (Chetty et al., 2014b). The authors find that teacher value added varies more in math than in ELA, but that future earnings are *less* related to math value added than to ELA value added. This suggests that math value added may partly reflect test prep that is not beneficial for future earnings, although there are other possible explanations for this pattern.

[7] We define $\tau(R_i)$ as a function of R_i because most U.S. states score exams such that students with the same raw scores receive the same scale scores. In a few states, scores are computed using the full vector of exam responses, $\{u_{iq}\}_{q=1}^{Q}$. See Appendix C.3 for further details on the scoring of exams.

exceeds a minimum threshold deemed to be proficient, $\underline{R}$. Proficiency levels are often the most important consideration in accountability systems. For example, NCLB tied school funding to growth in proficiency rates through its "adequate yearly progress" targets (Dee and Jacob, 2011).

We assume the teacher chooses e^* and α^* to maximize the expected proficiency rate in their class given costly effort. The teacher can only maximize *expected* proficiency because some aspects of students' performance are outside of their control (e.g., health on exam day). We assume the choice of the target level, α^*, is costless.[8]

We briefly characterize the teacher's optimal choice of α^* and e^* here; Appendix B.1 provides a full derivation. Intuitively, the teacher's optimal instruction level, α^*, is centered around students for whom additional skill is likely to make a difference in achieving proficiency. This choice balances the expected proficiency gains from moving α^* closer to the ability levels of some students with the expected losses from moving it further from other students' abilities (see Appendix Equation B6). The optimal value of α^* will typically be near the ability level of marginally-proficient students, but this value varies with the ability distribution in the class. This is analogous to the key insight in Neal and Schanzenbach (2010).

Our main focus is on the optimal effort level, e^*, which we characterize in Proposition 1:

Proposition 1. *The teacher's optimal level of effort, e^*, is increasing in:*

- *The number of students on the margin of expected proficiency; and*
- *The derivative of expected proficiency with respect to ability, $dE[\tau(R_i)|\theta_i]/d\theta_i$, for marginally-proficient students.*

Intuitively, teachers exert more effort when: 1) there are more students near the proficiency margin; and 2) when skill accumulation for these students is more likely to be rewarded by the exam in terms of achieving proficiency, i.e., when $dE[\tau(R_i)|\theta_i]/d\theta_i$ is larger. Put simply, effort is increasing in the *returns* to effort—as defined by its impact on the class proficiency rate (see Appendix Equation B7). [9]

The next subsection shows that the structure of a standardized exam impacts both of these factors.

3.2. **Exam design.** There are two main elements of exam design that affect the determinants of teacher effort in Proposition 1. First, test designers set a proficiency standard,

[8] The choice of α^* may be costly if teachers must change their lesson plans; we abstract from such costs.

[9] Macartney et al. (2018) show that teacher value added is positively related to the proportion of students in the classroom who are on the proficiency margin. The first part of our Proposition 1 mirrors this result, although we focus on how the proportion of marginal students varies with the exam structure.

which affects how many students are likely to be near the proficiency margin. In our framework, this standard corresponds to the choice of the raw score threshold, $\underline{R}$. In practice, test designers choose $\underline{R}$ based on the level of expected ability they deem to be proficient.[10] This choice determines where the proficiency margin falls in the distribution of ability. Exams in which 50 percent of students are deemed proficient will tend to have more marginally-proficient students than exams with a proficiency rate of 85 percent. A major criticism of NCLB was that states could set their own proficiency standards, which led to wide variation in proficiency rates (Reback et al., 2014).

Second, the derivative of expected proficiency, $dE[\tau(R_i)|\theta_i]/d\theta_i$, depends on how informative the exam questions are for a given level of ability, θ_i. To write an exam, test designers select topic areas and write questions to test concepts within each topic. These questions are then placed onto a unidimensional scale of ability, often using Item Response Theory (IRT). For example, the "three-factor" IRT model assumes that the probability of a correct answer to question q as is a logistic function of θ_i:

$$p_q(\theta_i) = c_q + \frac{1 - c_q}{1 + e^{-a_q(\theta_i - b_q)}}. \tag{4}$$

The parameter b_q is referred to as the question's *difficulty* because the probability of a correct answer is decreasing in b_q. Question difficulty is expressed in the same units as student ability, θ_i, and the derivative of the probability of a correct answer, $p'_q(\theta_i)$, is largest for students whose ability matches the difficulty level, $\theta_i = b_q$. The parameter a_q is known as the question's *discrimination*; higher values of a_q imply that the question is better able to discriminate between test takers with $\theta_i > b_q$ and $\theta_i < b_q$. Lastly, the c_q parameter defines the question's *guessability*; c_q is equal to the probability of a correct answer as $\theta_i \to -\infty$.[11]

The distribution of question parameters affects how precisely the exam measures ability at any given level of θ_i, which affects the magnitude of $dE[\tau(R_i)|\theta_i]/d\theta_i$. All else equal, $dE[\tau(R_i)|\theta_i]/d\theta_i$ is larger for marginally-proficient students when:[12]

- The exam has many questions in which the difficulty, b_q, is close to the level of marginally-proficient ability;
- The questions near the proficiency margin are more discriminating (higher a_q);
- The questions are less guessable (lower c_q);
- The exam has more questions overall (higher Q).

[10] Test designers often use the "bookmark method" to set the proficiency standard: the exam questions are displayed in order of difficulty, and educators are asked to bookmark the page at which a minimally-proficient student would stop providing correct answers.

[11] Test designers try out potential exam questions by adding them as un-scored items on other tests, and then use these responses to estimate the parameter values $\{a_q, b_q, c_q\}$.

[12] See Appendix B.2 for details on how these parameters affect $dE[\tau(R_i)|\theta_i]/d\theta_i$ near the proficiency margin.

Our data in Section 4 includes both proficiency standards and IRT parameters, and we show how these design elements affect test prep incentives.

3.3. **Incentives in math and ELA.** Below we show that math and ELA exams during this period often varied in structure, and thus created different incentives for test prep. Our framework does not explicitly model tradeoffs between subjects, but teachers would optimally spend more time on subjects with higher returns to test prep. Elementary school teachers face a direct tradeoff between math and ELA prep since they teach both subjects. Even in middle school, where teachers often teach one subject, the strength of incentives matters because test prep reduces time available for other material.

We make two assumptions in comparing incentives in math and ELA. First, our analysis implicitly assumes that the effect of test prep on student ability (equation 1) is the same in math and ELA. Many researchers hypothesize that school inputs matter more for learning in math than in ELA (Jackson et al., 2014). Our analysis does not speak to this hypothesis, and we believe that is it likely to play a role in the pattern of literature estimates. We ask instead whether test prep incentives can contribute to this pattern even if one does not assume differences in the production of math and ELA skills.

Second, we assume that teachers can observe the factors that affect test prep incentives (Proposition 1). Educators do not perfectly know the exam structure *ex ante*, but experienced teachers will know which students are likely to be near the proficiency margin. Further, states typically provide practice tests to help teachers prepare students for questions that are likely to appear on the exam. Research in the economics of education often finds that teacher effort responds to accountability and evaluation incentives (e.g., Taylor and Tyler, 2012; Imberman and Lovenheim, 2015; Aucejo et al., 2020). There is also a large literature in education that shows how accountability alters teachers' instruction choices (e.g., Koretz et al., 1996; Pedulla et al., 2003; Jennings and Bearak, 2014). Thus we would find it surprising if most teachers were unaware of incentives created by the design of exams.

4. Data collection and analysis

This section describes our data, methods, and main findings on the incentives for test prep in math and ELA exams.

4.1. **Data and methods.** We collected data from the technical reports of grade 3–8 math and ELA exams in six U.S. states that are the setting for most of the papers in Table 1: Florida, Illinois, Massachusetts, New York, North Carolina, and Texas. We obtained reports from two years for each state: one year in 2000–2003 (pre-NCLB), and another in 2006–2008 (NCLB era). In each time period, we used the earliest year for which we could find a report

with all the information necessary for our analysis. Appendix Table C1 shows our data sources and the exam years for each state.

Our analysis relies on three types of data from these reports. First, we use information on the questions that appeared on the exams in each subject and grade. Some reports provide the IRT parameters $\{a_q, b_q, c_q\}$ for every exam question; others report only the distribution of these parameters. Second, we use information on how raw scores (total correct answers) map into scale scores. Third, the reports provide data on the realized distribution of scale scores from the exam administration, and on the minimum score necessary for proficiency.

We use this data to simulate test scores for the population of students who took each exam (defined by a state, year, grade, and subject). Our simulations for each exam proceed as follows:

(1) Create $i = 1, \ldots, 1000$ test takers at each ability level $\theta \in \{-5, -4.9, \ldots, 0, \ldots, 4.9, 5\}$, where we begin with the prior that $\theta_i \sim N(0, 1)$.
(2) Use the question parameter data to draw a random vector of exam responses, $\{u_{i1}, \ldots, u_{iQ}\}$, based on each test taker's ability, θ_i, and the IRT model (equation 4).[13] Each u_{iq} indicates whether test taker i answered question q correctly.
(3) Compute each test taker's raw score, $R_i = \sum_{i=1}^{Q} u_{iq}$.
(4) Convert raw scores to scale scores using the scaling data.
(5) Update the distribution of ability, θ_i, to match the observed scale score distribution in the population of students that took the exam.[14]

Appendix C provides details on our data and simulations. We also provide our simulation codes as supplementary material for this paper.

From this simulated data we compute the key statistics that relate to Proposition 1: the density of the ability distribution and the derivative of expected proficiency for marginally-proficient students. Throughout our analysis, we define "marginally-proficient" as the level of ability at which the likelihood of proficiency is 50 percent, and we denote this ability level by $\underline{\theta}$. We also show statistics on the exam structure, which come directly from the technical reports.

Our use of simulated data comes with several caveats. First, our simulated data is designed to match the observed score distribution, which includes any impacts of real-world test prep. This can affect our estimates of the density of the ability distribution at $\underline{\theta}$ because test

[13] When we do not have question-level data on the IRT parameters $\{a_q, b_q, c_q\}$, we draw questions randomly from a normal distribution with mean/variance equal to the values from the report.
[14] The prior that $\theta_i \sim N(0, 1)$ is the same assumption that test designers make in scoring exams. We then update the density of θ_i so that the distribution of our simulated scale scores matches the distribution of scale scores from the exam administration. See Appendix C.5 for details.

prep can alter this distribution.[15] Thus our results implicitly assume that real-world test prep does not significantly alter the density of the ability distribution at $\underline{\theta}$. We think this is a reasonable assumption since most of the literature estimates in Table 1 are modest in magnitude. Second, our data contain additional measurement error from simulation. But we use a large number of simulated test takers, so the magnitude of this error is small. Appendix Table A6 shows that our main results do not change significantly when we re-run our simulations.

4.2. **Example.** Figure 1 illustrates how exam design can affect test prep incentives using New York's 2006 8$^{\text{th}}$ grade math and ELA tests. Panel A plots the cumulative distribution functions of question difficulty, b_q, for the math (red circles) and ELA (black triangles) exams.[16] Each marker represents one exam question. Question difficulty is expressed in units of individual ability, θ_i, which is scaled to be mean 0 and standard deviation 1 in the reference population for each exam. The vertical lines denote the level of ability at which the probability of achieving proficiency is 50 percent, $\underline{\theta}$. Marker sizes are proportional to the question's discrimination parameter, a_q.

Three features of New York's 8$^{\text{th}}$ grade math exam made it a more precise measure of ability for marginally-proficient students than its ELA exam. First, the math exam featured many questions in which the difficulty, b_q, was close to proficiency margin, $\underline{\theta}$, whereas the ELA exam had fewer such questions. Second, the math exam had more questions, Q, than the ELA exam (68 vs. 39); this gave test takers more opportunity to demonstrate their ability. Finally, the math questions were more discriminating on average than the ELA questions, as measured by the mean discrimination parameter, $\bar{a}_q$ (1.23 vs. 1.02).

These differences meant that the derivative of expected proficiency for marginally-proficient students was much larger in math. Panel B of Figure 1 plots expected proficiency in math (red circles) and ELA (black triangles) at 0.1 increments of ability. At 50 percent expected proficiency, the derivative with respect to θ_i is 2.6 for the math exam. This means that a 0.1 unit increase in a student's ability would increase their likelihood of achieving proficiency by 26 percentage points. For the ELA exam, the derivative at $\theta_i = \underline{\theta}$ was only 1.4.

Test prep incentives also depend on the proficiency standard, which affects the number of students who are near the proficiency margin. Panel B illustrates this with vertical bars, which show the ability distributions for each exam. New York's 8$^{\text{th}}$ grade exams had proficiency rates of 54 percent in math and 49 percent in ELA, so marginal students were in

[15] Specifically, our simulated data gives us the density of post-prep ability, θ_i, while our framework is based on the density of pre-prep ability, α_i. The other key statistic in Proposition 1—the derivative of expected proficiency at $\underline{\theta}$—is not affected by real-world test prep, as it depends only on the design of the exam.

[16] Panel A does not plot a few very easy ELA questions ($b_q < -2$) to make the graph more readable.

the middle of the ability distribution for both exams. But the fraction of students near the proficiency margin varied widely across exams in our sample, as we show below.

4.3. **Summary statistics for all test takers.** Table 2 presents summary statistics for the exams in our sample. Our sample contains 122 exams, which come from six states (FL, IL, MA, NY, NC, TX), two years (pre-NCLB and NCLB era), two subjects (math and ELA), and up to six grade levels.[17] Panel A reports averages over all 122 exams. Panels B–C present statistics for each state in the pre-NCLB and NCLB years. Within each panel, we show averages computed over the state's grade 3–5 and grade 6–8 exams.

The statistics in Table 2 correspond to the average student who took each exam (computed in our simulations). We discuss results for the average student before turning to marginally-proficient students so that we can highlight overall differences between math and ELA exams. For the average test taker, Table 2 shows the likelihood of achieving proficiency (columns A–B), the proportion of correct answers (columns C–D), and the *derivative* of the proportion of correct answers with respect to ability, θ_i. (columns E–F).

We find that math and ELA exams differed in each of these three statistics. First, math exams tended to have lower proficiency rates than ELA exams (columns A–B). Across all states, grades, and time periods, 66 percent of math test takers achieved proficiency, as compared with 70 percent of ELA test takers (Panel A). This difference was most pronounced for grade 6–8 exams in the NCLB era, for which math proficiency was 10 percentage points lower than ELA proficiency (0.61 vs. 0.71). The difference in proficiency rates was modest in other grades and time periods, but there was wide variation across states (e.g., Massachusetts vs. Texas) and over time (e.g., in Illinois). This shows that test designers tended to set a "higher bar" for achieving proficiency in math exams than in ELA exams (relative to the ability of the average test taker).

Second, math exams were systematically harder for the average test taker (columns C–D). Across all exams, math test takers got 63 percent of the questions correct, as compared with 70 percent correct in ELA (Panel A). This difference is again more pronounced in grades 6–8, for which the proportion correct was roughly 10 percentage points lower in math than in ELA in both time periods. Thus test designers tended to write more difficult questions in math than in ELA, particularly at higher grade levels.[18]

Finally, the derivative of the probability of a correct answer was larger for the average test taker in math than in ELA (columns E–F). This statistic is the mean value of the derivative of equation (4) with respect to θ_i, or $p'_q(\theta_i)$. The average value of this derivative was 0.18 in math and 0.16 in ELA (Panel A); these values imply that a 0.1 unit increase

[17] In the pre-NCLB years, some states administered tests in only a few grades.

[18] We emphasize that the proficiency standard and question difficulty are two distinct exam design choices. For example, a test could contain only easy questions, but require a near perfect score to achieve proficiency.

in a test taker's ability would increase the proportion of correct answers by 1.8pp in math and 1.6pp in ELA. This variation is attributable to differences in the difficulty of math and ELA exams (columns C–D); question difficulty was typically better-aligned with the ability of the average test taker in math than in ELA.

These differences between math and ELA tests created different incentives for educators to prepare marginally-proficient students for exams. We now turn to these results.

4.4. **Main results.** Table 3 presents our main results on the factors that affect test prep incentives as described in Proposition 1. This table has the same structure as Table 2, but we present statistics from our simulations for test takers on the proficiency margin, i.e., those with ability $\theta_i = \underline{\theta}$. Columns (A)–(B) show the density of the ability distribution at $\underline{\theta}$, and columns (C)–(D) show the derivative of expected proficiency, $dE[\tau(R_i)|\theta_i]/d\theta_i$, evaluated at $\theta_i = \underline{\theta}$. These correspond to the two factors that affect the teacher's optimal level of test prep effort in Proposition 1. Columns (E)–(F) in Table 3 show the product of these two terms, and column (G) shows the math/ELA ratio of these products.

Our main finding is that the incentives for test prep were typically stronger in math than in ELA. Both of the factors in Proposition 1 contributed to this pattern.

First, there were more students near the proficiency margin in math than in ELA (columns A–B). Averaged across all exams in our sample, the density of the ability distribution near the proficiency margin was 0.32 in math and 0.30 in ELA. These density values imply that the number of marginally-proficient students was roughly seven percent greater in math than in ELA on average ($0.32/0.30 \approx 1.07$). The math/ELA difference in density values is most pronounced for grade 6–8 exams in the NCLB era (0.34 vs. 0.30). This pattern arises because the bar for achieving proficiency was typically closer to the middle of the statewide ability distribution in math than in ELA (see Table 2).

Second, the derivative of expected proficiency for marginally-proficient students was larger in math than in ELA. Averaged across all exams, this derivative is 1.50 in math and 1.32 in ELA. This implies that a 0.1 unit increase in a marginally-proficient test taker's ability would increase their likelihood of achieving proficiency by 15pp in math, and 13pp in ELA. This pattern is systematic across the exams in our sample; the derivative of expected proficiency is larger in math for nearly all exam groups in Table 3. This is partly due to the fact that math exams were more difficult than ELA exams (Table 2), as we discuss in more detail in Section 4.5.

Thus the product of the two factors that affect test prep incentives is systematically higher in math than in ELA. On average, the product of the density of the ability distribution and the derivative of expected proficiency for marginally-proficient students was 0.49 in math, and 0.40 in ELA (columns E–F). The ratio of these two products is 1.22, suggesting that

incentives for test prep as defined by Proposition 1 were 22 percent higher in math. The math/ELA ratio of these products is the largest for grade 6–8 exams in the NCLB era (Panel C), but this ratio is above one for the vast majority of exam groups in our sample.

Table 4 shows that these math/ELA differences are statistically significant at conventional levels. In this table, we regress the statistics from our simulations on an indicator for the math exam and state × year × grade fixed effects. Panel A uses dependent variables that correspond to the summary statistics for all exam takers from Table 2. Panel B uses the statistics for marginally-proficient test takers from Table 3 as outcome variables. Column (A) shows the mean value of each statistic for ELA exams, and column (B) displays the coefficient on the math indicator in a regression with all 122 exams. All of the math/ELA differences discussed above are significant at $p < 0.05$. The last row of Panel B shows that our main result—the product of the test prep factors—is significantly larger in math than in ELA at $p < 0.01$. Columns (C)–(F) in Table 4 report regression results separately by grade group (3–5 vs. 6–8) and time period (pre-NCLB vs. NCLB era). The math/ELA differences continue to be statistically significant in most cases despite the smaller number of exams.

4.5. **Exam design components.** The most systematic pattern in Table 3 is that the derivative of expected proficiency is larger in math than in ELA near the proficiency margin. This subsection shows that this pattern can be explained by differences in the structure of math and ELA exams.

Figure 2 displays four elements of exam structure for our sample of exams. Each panel plots exam design parameters in math (y-axis) against parameters for the ELA exam (x-axis) in the same state, year, and grade group. The marker text indicates the state and year (e.g., MA-06), and the color indicates the grades (blue = 3–5; grey = 6–8). Panel A shows how the difficulty of the exam questions aligned with the ability of marginally-proficient test takers, as measured by the mean squared error (MSE) between b_q and $\underline{\theta}$.[19] Panels B–D show the number of questions, Q, the mean discrimination parameter, a_q, and the mean guessability parameter, c_q. Each panel includes a 45 degree line to illustrate whether parameters tend to be higher in math or ELA.

There are significant differences between math and ELA exams in each of these design components. Specifically, we find that:

- Panel A. Math question difficulty was typically closer to the level of marginally-proficient ability than ELA question difficulty, with an average difference in MSE of 0.28 (in standardized units of ability).

[19] As an example, the blue MA-06 marker in Panel A shows that the MSE between b_q and $\underline{\theta}$ for the 2006 grade 3–5 exams in Massachusetts was 1.9 in math, and 2.6 in ELA. This marker is below the 45 degree line, implying that question difficulty in the ELA exam was less aligned with the ability of marginally-proficient test takers than in the math exam.

- Panel B. Math exams featured 8 more questions than ELA exams on average.
- Panel C. Math questions were more discriminating than ELA questions on average, with a mean gap in a_q of 0.06.
- Panel D. Math questions were about 2 percentage points less "guessable" than ELA questions on average.

Table 5 shows that each of these factors are related to a larger derivative of expected proficiency for marginally-proficient students. This table presents regression results where the dependent variable is the derivative of expected proficiency at $\theta_i = \underline{\theta}$, and the covariates are the four exam design components from Figure 2. We standardize each covariate to be mean zero and SD one in our full sample of exams, so coefficients represent the effect of a one standard deviation increase in each parameter. We find that the derivative of expected proficiency at $\theta_i = \underline{\theta}$ is increasing in mean question discrimination, a_q, and the total number of questions, Q. This derivative is decreasing in mean question guessability, c_q, and the MSE between b_q and $\underline{\theta}$.[20] Taken together, the results from Figure 2 and Table 5 show that all four design components contributed to the larger derivative of expected proficiency in math, and the contribution of each component was similar in magnitude (column F of Table 5).

4.6. **Sensitivity checks.** Our main results are robust to using a subset of our sample with the highest-quality data on exam design. Our full sample includes some exams for which we only observe the *distribution* of IRT parameters, rather than the IRT parameters for each individual question (see Appendix Table C2). In other cases we observe p-values (proportion of correct answers) for each question rather than the IRT parameters. Appendix Table A3 shows that we continue to find stronger test prep incentives in math relative to ELA if we focus only on exams with question-level IRT data.

Our findings are also robust to an alternate method of weighting the distribution of test taker ability, θ_i. For our main analysis, we re-weight the distribution of θ_i to match the statewide distribution of scores in the year that the exam was administered. As a robustness check, we do not re-weight, and instead assume ability follows a standard normal distribution, i.e., $\theta_i \sim N(0, 1)$. This is the assumption that test designers make when they set the scale for an exam. Thus, this robustness check shows how our findings would change if the exams in our sample were given to their *reference* populations rather than to the students who actually took the exams.[21] Appendix Table A3 shows that this alternate weighting method does not change our finding that test prep incentives are stronger in math than in ELA.

As a final sensitivity check, we show how our results vary with the definition of the "proficiency margin." Our benchmark results are for the level of ability, $\underline{\theta}$, at which the

[20] Appendix Table A2 shows that the derivative of expected proficiency at $\theta_i = \underline{\theta}$ has a concave relationship with the number of questions, Q, and with the MSE between b_q and $\underline{\theta}$.

[21] The reference population is often the set of test takers in the first year that an exam was administered.

likelihood of achieving proficiency is 50 percent. Table 6 shows how these results change when we use a wider bandwidth around $\underline{\theta}$. Specifically, we compute the density of the ability distribution, the derivative of expected proficiency, and the product of these two terms at $\underline{\theta} \pm h$, where h takes the values 0.2, 0.4, 0.6, and 0.8. We compute the average of each statistic within these bandwidths, and use these averages as the dependent variables in Table 6. Column (B) replicates our benchmark results from Panel B of Table 4, and columns (C)–(F) show results with different bandwidths. At the bottom of Table 6 we report the expected proficiency rate at the low and high end of each ability range ($\underline{\theta} - h$ and $\underline{\theta} + h$).

The results in Table 6 show that the derivative of expected proficiency is larger in math than in ELA over a significant range around $\underline{\theta}$, but this difference disappears at wide bandwidths. Column (C) shows that our main results are similar when we include ability levels within $h = 0.2$ units of the proficiency margin, $\underline{\theta}$; this includes test takers with expected proficiency ranging from 30–70 percent. The math/ELA difference in the derivative of expected proficiency fades as we widen the bandwidth, and it is equal to zero at a bandwidth of $h = 0.8$ (column F). The bandwidth of $\underline{\theta} \pm 0.8$ is very wide in the sense that it includes test takers with expected proficiency ranging from 1–99 percent. This convergence arises because, relative to math exams, ELA exams typically measured ability more precisely for students who were *far* from the proficiency margin.[22]

We do not view the evidence in Table 6 as indicating that our results are not robust; rather, it illustrates that our conclusions rely on the assumption that teachers focus their test prep on marginally-proficient students. Our hypothesis does not hold if teachers' test prep has significant benefits for students who are well above or well below the proficiency margin, as ELA exams measure learning more precisely for these students. There is compelling evidence that teachers adjust their practices and level of instruction in response to incentives (Neal and Schanzenbach, 2010; Duflo et al., 2011; Kane et al., 2011). Thus we believe that differences in the design of math and ELA exams are likely to affect teacher behavior, particularly when they are held accountable by policies that emphasize proficiency rates.

4.7. **Why are math exams more discriminating at the proficiency margin?** Our results show that math exams were systematically more informative for ability near the proficiency margin than ELA exams during 2000–2008. We think this is likely to be a general phenomenon, for several reasons.

It is often said that math requires cumulative knowledge, i.e., that future learning depends critically upon past learning.[23] This would lead math questions to be more discriminating on

[22] The main reason that ELA exams measure ability more precisely far from the proficiency margin is illustrated in Panel A of Figure 2: the SD of question difficulty tends to be higher in ELA than in math.

[23] As the psychologist Steven Pinker put it: "Mathematics is ruthlessly cumulative, all the way back to counting to ten" (Pinker, 1997).

a unidimensional scale of ability (as measured by the a_q parameter). Specifically, this adage implies that the likelihood that a low-ability test taker correctly answers a hard question is lower in math than in ELA. One might also argue that math questions are less subjective than ELA questions, which may make them harder for low-ability test takers to guess (as measured by the c_q parameter).

The nature of learning in math may make it easier to write an exam that is a precise measure of proficiency. The ideal exam is highly informative for ability in the region that educators care most about. Exam designers try out potential questions and eliminate those that are too hard, too easy, or not discriminating enough, but this takes time and resources (Chingos, 2012; Topol et al., 2012). If math requires cumulative learning, there may be more agreement on what it means to be proficient. It is easier to write exam questions that are informative about student proficiency when there is more agreement on the standards (Bergman et al., 2021). There is also a longstanding concern about U.S. students' underperformance in mathematics (Goodman, 2019). This may explain why test designers tend to set higher proficiency standards in math than in ELA (relative to the distribution of student ability).

Lastly, math questions can usually be described briefly, while ELA exams often feature reading passages. Thus ELA tests must allow time for students to read before they start answering questions. This fact would allow math exams to feature more questions in a fixed block of time, as we found in Panel B of Figure 2.

Test designers could modify some of these design features to equalize the test prep incentives in math and ELA exams (if they wished to do so). For example, test designers could add extra questions to ELA exams and allow for more time in the exam administration. They could also devote more effort to screening potential ELA questions based on their difficulty and discrimination.

5. Relationship between test prep incentives and literature estimates

This section discusses the implications of our findings in Section 4 for the literature patterns that we documented in Section 2.

5.1. **Theoretical relationship.** Under the assumptions of our framework, differences in the design of math and ELA exams can partly explain why math effects tend to be larger in the literatures in Table 1. Research on accountability policies measures the effects of an increase in the stakes of an exam, which gives teachers an incentive to engage in more test prep. Proposition 1 predicts that the increase in test prep effort would be larger in math than in ELA because there are greater returns in terms of proficiency. Similarly, good schools or good teachers may be better at converting effort into test score gains, or they may have

lower costs of effort. If these actors choose effort as in our framework, the effects of admission to a good charter school would be larger in math than in ELA, and the variation in teacher value added would be larger in math. Appendix B.4 provides a more formal discussion of these arguments.

A simple comparison of magnitudes suggests that test prep incentives can account for some portion—but not all—of the variation in math and ELA effects in the literature. The math estimates in the literature are roughly 50 percent larger than the ELA estimates on average (Table 1), while the mean difference in test prep incentives is 22 percent (Table 3). The precise connection between these statistics depends on many features of the education production function, which we do not quantify in this paper. While simply suggestive, we think these magnitudes present a compelling case for test prep incentives as a plausible explanation for the pattern in the literature.

5.2. **Empirical relationship.** To explore this relationship empirically, we link the results of our simulations to the estimates from our literature review. Specifically, we begin with the math/ELA ratio of literature estimates from column (H) of Table 1. In 19 cases, the authors' data comes from one of the states in our sample of exams. We link these 19 literature estimates to our simulation results using the grades and time periods (pre-NCLB and NCLB era) that match the authors' sample. For example, we link our results for the Massachusetts exams in 2000 (grade 8) and 2006 (grades 6–8) to Abdulkadiroğlu et al. (2011)'s estimates for Boston charter school students in 2002–2008 (grades 6–8).[24] Our explanatory variable of interest is the math/ELA ratio of test prep incentives for marginally-proficient students, as in column (G) of Table 3. We compute the average of this ratio over all exams that link to the authors' sample, and relate this to the math/ELA ratio of the authors' estimates.

Figure 3 shows that there is a strong positive relationship between the math/ELA ratios of test prep incentives and literature estimates. The y-axis in this figure is the math/ELA ratio of the authors' estimates, and the x-axis is the mean math/ELA ratio of test prep incentives from our simulations. The text of each marker indicates the paper, with parentheses indicating the grades included in the authors' samples. The dashed line shows that these variables are positively related with a slope of approximately 0.8. This implies that a 10 percent increase in the math/ELA ratio of test prep incentives is associated with an eight percent increase in the math/ELA ratio of literature estimates.

[24] To increase the number of matches, we do not require that the exam years are exactly the same as the authors' sample years, but only that they come from the same pre-NCLB and/or NCLB period. This implicitly assumes that the exam structure is similar over time, which is an important objective for test designers.

Table 7 presents results from regressions that are analogous to Figure 3. This table displays OLS coefficients from bivariate regressions of the ratio of literature estimates (y-axis in Figure 3) on the ratio of test prep incentives (x-axis in Figure 3). Column (A) includes all 19 matched estimates in the regression, and columns (B)–(D) present results separately for the accountability, charter school, and teacher value added papers. To reduce noise from outliers, we winsorize the dependent variable by setting the maximum/minimum literature estimates to their 90th/10th percentile values.[25] Panel A shows our preferred specification, which weights each literature estimate by the number of matched exams from our sample. This specification gives more weight to observations for which we have more data to estimate the strength of test prep incentives. Panel B presents results that instead give equal weight to each literature estimate, regardless of the number of matched exams.[26]

The OLS relationship between the math/ELA ratios of test prep incentives and literature estimates is 0.83 in our preferred specification, and it is statistically significant at $p < 0.01$ (column A). This relationship is strongest in the five papers in the accountability literature, with a coefficient above two (column B). There is also a strong relationship between these variables in the charter school papers (column C), although slightly smaller in magnitude. The relationship is weaker but still positive in the teacher value added literature (column D), consistent with the discussion in Section 2.

In sum, Figure 3 and Table 7 provide evidence that variation in test prep incentives can explain variation in math and ELA estimates from the literature. This relationship is large in magnitude and statistically significant despite the small number of papers. Both of these facts suggest that differential incentives for test prep are an important mechanism in papers where accountability is important.

5.3. **Generalizability to other states.** We next ask whether our findings generalize to states that we did not include in our main analysis. Figure 4 displays data on the structure of standardized exams in other states during the NCLB era. Panel A shows the proficiency rate in each state averaged across all grade 4 and 8 test takers in 2007; this information comes from the National Assessment of Educational Progress (NAEP) and covers nearly all 50 states. Panels B–D present data from the technical reports that we were able to find for exams in the early NCLB era (2006–2009).[27] These panels show the proportion of correct

[25] This changes the math/ELA ratio of Dobbie and Fryer Jr (2011)'s grade 6–8 estimates from 4.87 to 2, and it changes the ratio of Hoxby and Rockoff (2004)'s estimates from 0.47 to 1.05. We find qualitatively similar results without winsorizing or dropping these two studies, but the estimates are noisier (see Appendix Tables A4–A5). In Figure 3, we moved Dobbie and Fryer Jr (2011)'s grade 6–8 estimate from 4.87 to 2.5 to improve readability.

[26] In both panels, we cluster standard errors at the literature estimate level.

[27] Education data from states that are not in our main sample are generally hard to obtain (which is why there are few papers on these states in the literature). Appendix Table C1 shows the states and years for which we could find technical reports.

answers averaged across all grades (Panel B), the mean number of questions (Panel C), and the mean discrimination parameter, a_q (Panel D). We show values for the six states in our main sample in blue, and values for other states in grey. Figure 4 otherwise has the same structure as Figure 2, with math values on the y-axes, and ELA values on the x-axes.

Figure 4 shows that the patterns of exam structure that we identified in our main sample also hold in other states. Specifically, we find that:

- Panel A. Math exams had lower proficiency rates than ELA exams on average (0.67 vs. 0.72).
- Panel B. Math exams were more difficult than ELA exams, as measured by the mean proportion of correct answers (0.61 vs. 0.65).
- Panel C. Math exams contained more questions than ELA exams on average (63 vs. 57).
- Panel D. Math questions were more discriminating than ELA questions, as measured by the mean value of the a_q parameter (0.94 vs. 0.85).

These patterns are the same as those in Table 2 and Figure 2 above. Further, Figure 4 shows that the states in our sample are not outliers relative to other states. This suggests that our finding of stronger test prep incentives in math is not unique to the states in our analysis.

If our findings generalize to other states, then test prep may also be an important mechanism in national studies of NCLB. Our literature review included two papers that examined the effects of NCLB, and both found significantly higher effects in math than in ELA (Wong et al., 2009; Dee and Jacob, 2011). The patterns in Figure 4 suggest that differences between math and ELA in test prep incentives may partly contribute to this finding. A caveat is that the test prep incentives under NCLB were complex because schools had to achieve proficiency in different subgroups of students. NCLB also included safe harbor provisions for schools that did not meet the main proficiency standards, and these provisions varied both across states and subjects (Ahn and Vigdor, 2014).

5.4. **Math/ELA estimates in research on class size.** As further evidence that accountability pressures contribute to the pattern of larger math effects, we note that this pattern does *not* arise in an older literature on the impacts of class size on student achievement.[28] Many papers in this literature use U.S. data from the 1980s or earlier, before the adoption of consequential state accountability policies (Dee and Jacob, 2011). In a meta-analysis of 80 studies conducted between 1895 and 1978, Glass and Smith (1979) find that instruction

[28] We are grateful to the editor for this insightful observation.

subject (math, reading, etc.) was not a significant predictor of the relationship between class size and achievement.[29]

Appendix Table A7 shows that math and ELA estimates tend to be similar in the most prominent and well-identified papers in the class size literature.[30] Across six class size papers that use data from settings with low accountability, we find that the average ratio of math/ELA point estimates is 0.89. This includes Krueger (1999)'s analysis of the Tennessee Student/Teacher Achievement Ratio (STAR) experiment, which finds, if anything, smaller gains in math than in reading from class size reductions.[31] In international settings where standardized exams were administered without much accountability—Israel and Bolivia—Angrist and Lavy (1999) and Urquiola (2006) find broadly similar effects in math and language arts, depending on the specification. Notably, Appendix Table A7 also includes three papers that use data from the 1990s in U.S. states *after* the implementation of an accountability policy: Wisconsin (Molnar et al., 1999), Texas (Rivkin et al., 2005), and California (Jepsen and Rivkin, 2009). Each of these papers finds larger point estimates in math than in ELA, with an average math/ELA ratio of 1.34.

Although not conclusive, the findings from class size research suggest that accountability pressures are an important driver of variation in math and ELA estimates in other literatures. If the pattern of larger math effects was *only* driven by differences in the education production function for each subject, one would expect this pattern to also appear in class size research from years with low accountability.

6. Conclusion

This paper showed that the design of standardized accountability exams affects incentives for educators to engage in test prep. Using data on grade 3–8 state exams in 2000–2008, we showed that math exams often had more students near the proficiency margin than ELA exams, and they typically measured ability more precisely in this region. This created an incentive for teachers to spend more time on math prep than on ELA prep, consistent with the common finding of larger math effects in literatures where accountability matters. Our results suggest that accountability pressures and test prep incentives are an important driver of this pattern, although other factors likely play a role as well.

[29] Glass and Smith do not report separate estimates for math and reading, but they write: "Among those factors of discrimination that produced virtually identical regression lines were ... 'subject taught' " (p. 12).
[30] There is disagreement on how to interpret the evidence from this literature (Hanushek, 1997; Krueger, 2003), but many well-identified papers find that class size reductions increase achievement (Schanzenbach, 2014).
[31] Hoxby (2000) uses Connecticut data from 1986–1998 and employs two different identification strategies to estimate class size effects. Hoxby's estimates do not vary systematically between math and ELA, although she characterizes the results as precisely-estimated zeroes.

We think our findings are attributable to several inherent features of math that make it an easier subject to test, and thus are likely to affect teacher behavior. Yet test prep incentives vary widely across settings, and standardized exams are frequently redesigned. In the 2010s, for example, the Common Core movement led several states to switch to the PARCC assessments, which were notoriously long and difficult (Jochim and McGuinn, 2016). More recently, many states have adopted computer adaptive testing, which increases precision by tailoring question difficulty to each student (Martin and Lazendic, 2018).

Thus we advise researchers to provide information on accountability policies and exam structure when they use test scores as outcome variables. It is important to know which exam metrics are used for accountability and the number of students who are likely to be on the margin of achieving these metrics. In the case of a binary proficiency standard, the proficiency rate is a simple and informative statistic that describes where marginal test takers fall in the distribution of ability. Researchers should also report features of exam design that affect how precisely the exam measures ability, such as the average difficulty and number of questions.[32] This can shed light on whether variation in estimates across subjects may be driven by test prep, and whether effects are likely to generalize to settings with different exams and accountability pressures.

[32] Many technical reports also include the exam's Standard Error of Measurement, which shows how precision varies across the range of scores.

References

Abdulkadiroğlu, A., J. D. Angrist, S. M. Dynarski, T. J. Kane, and P. A. Pathak (2011). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *The Quarterly Journal of Economics 126*(2), 699–748.

Ahn, T. and J. Vigdor (2014). The impact of No Child Left Behind's accountability sanctions on school performance: Regression discontinuity evidence from North Carolina. NBER Working Paper No. 20511.

Akerhielm, K. (1995). Does class size matter? *Economics of Education Review 14*(3), 229–241.

Angrist, J. D. and V. Lavy (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly journal of economics 114*(2), 533–575.

Aucejo, E., T. Romano, and E. S. Taylor (2020). Does evaluation change teacher effort and performance? quasi-experimental evidence from a policy of retesting students. *The Review of Economics and Statistics*, 1–45.

Bacher-Hicks, A., T. J. Kane, and D. O. Staiger (2014). Validating teacher effect estimates using changes in teacher assignments in Los Angeles. Technical report, National Bureau of Economic Research.

Bergbauer, A., E. A. Hanushek, and L. Woessmann (2018). Testing. NBER Working Paper No. 24836.

Bergman, P., E. Kopko, and J. E. Rodriguez (2021). Using predictive analytics to track students: Evidence from a seven-college experiment. NBER Working Paper No. 28948.

Chay, K. Y., P. J. McEwan, and M. Urquiola (2005). The central role of noise in evaluating interventions that use test scores to rank schools. *American Economic Review 95*(4), 1237–1258.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *The American Economic Review 104*(9), 2593–2632.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review 104*(9), 2633–2679.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics 93*(9-10), 1045–1057.

Chingos, M. M. (2012). Strength in numbers: State spending on K-12 assessment systems. Technical report, Brown Center on Education Policy at Brookings.

Cohodes, S. R., E. M. Setren, and C. R. Walters (2021). Can successful schools replicate? scaling up boston's charter school sector. *American Economic Journal: Economic Policy 13*(1), 138–67.

Condie, S., L. Lefgren, and D. Sims (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review 40*, 76–92.

Corcoran, S. P., J. L. Jennings, and A. A. Beveridge (2011). Teacher effectiveness on high-and low-stakes tests. Society for Research on Educational Effectiveness.

Craig, S. G., S. A. Imberman, and A. Perdue (2013). Does it pay to get an A? school resource allocations in response to accountability ratings. *Journal of Urban Economics 73*(1), 30–42.

Curto, V. E. and R. G. Fryer Jr (2014). The potential of urban boarding schools for the poor: Evidence from seed. *Journal of Labor Economics 32*(1), 65–93.

Dee, T. S. and B. Jacob (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management 30*(3), 418–446.

Dobbie, W. and R. G. Fryer Jr (2011). Are high-quality schools enough to increase achievement among the poor? evidence from the harlem children's zone. *American Economic Journal: Applied Economics 3*(3), 158–87.

Duflo, E., P. Dupas, and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review 101*(5), 1739–1774.

Epple, D., R. Romano, and R. Zimmer (2016). Charter schools: A survey of research on their characteristics and effectiveness. In *Handbook of the Economics of Education*, Volume 5, pp. 139–208. Elsevier.

Figlio, D. and S. Loeb (2011). School accountability. In *Handbook of the Economics of Education*, Volume 3, pp. 383–421. Elsevier.

Figlio, D. N. and C. E. Rouse (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics 90*(1-2), 239–255.

Figlio, D. N. and J. Winicki (2005). Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics 89*(2-3), 381–394.

Fuller, B., J. Wright, K. Gesicki, and E. Kang (2007). Gauging growth: How to judge No Child Left Behind? *Educational Researcher 36*(5), 268–278.

Glass, G. V. and M. L. Smith (1979). Meta-analysis of research on class size and achievement. *Educational evaluation and policy analysis 1*(1), 2–16.

Gleason, P., M. Clark, C. C. Tuttle, and E. Dwoyer (2010). The evaluation of charter school impacts: Final report. ncee 2010-4029. Technical report, National Center for Education Evaluation and Regional Assistance.

Goldhaber, D., J. Cowan, and J. Walch (2013). Is a good elementary teacher always good? assessing teacher performance estimates across subjects. *Economics of Education Review 36*, 216–228.

Goodman, J. (2019). The labor of division: Returns to compulsory high school math coursework. *Journal of Labor Economics 37*(4), 1141–1182.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis 19*(2), 141–164.

Hemelt, S. W. (2011). Performance effects of failure to make Adequate Yearly Progress (AYP): Evidence from a regression discontinuity framework. *Economics of Education Review 30*(4), 702–723.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher 37*(6), 351–360.

Holmstrom, B. and P. Milgrom (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics and Organization 7*(1), 24–52.

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics 115*(4), 1239–1285.

Hoxby, C. M., J. Kang, and S. Murarka (2009). Technical report: How New York City charter schools affect achievement. *New York City Charter Schools Evaluation Project.*

Retrieved July 5, 2013.

Hoxby, C. M. and J. E. Rockoff (2004). *The impact of charter schools on student achievement*. Department of Economics, Harvard University Cambridge, MA.

Imberman, S. A. and M. F. Lovenheim (2015). Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system. *Review of Economics and Statistics 97*(2), 364–386.

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy 126*(5), 2072–2107.

Jackson, C. K., J. E. Rockoff, and D. O. Staiger (2014). Teacher effects and teacher-related policies. *Annu. Rev. Econ. 6*(1), 801–825.

Jacob, B. and J. Rothstein (2016). The measurement of student ability in modern assessment systems. *Journal of Economic Perspectives 30*(3), 85–108.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics 89*(5-6), 761–796.

Jacob, B. A. and L. Lefgren (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of Labor Economics 26*(1), 101–136.

Jacob, B. A. and S. D. Levitt (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics 118*(3), 843–877.

Jennings, J. L. and J. M. Bearak (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher 43*(8), 381–389.

Jepsen, C. and S. Rivkin (2009). Class size reduction and student achievement the potential tradeoff between teacher quality and class size. *Journal of human resources 44*(1), 223–250.

Jochim, A. and P. McGuinn (2016). The politics of the Common Core assessments: Why states are quitting the PARCC and Smarter Balanced testing consortia. *Education Next 16*(4), 44–53.

Jones, M. G., B. D. Jones, B. Hardin, L. Chapman, T. Yarbrough, and M. Davis (1999). The impact of high-stakes testing on teachers and students in North Carolina. *The Phi Delta Kappan 81*(3), 199–203.

Kane, T. J., J. E. Rockoff, and D. O. Staiger (2008). What does certification tell us about teacher effectiveness? evidence from New York City. *Economics of Education review 27*(6), 615–631.

Kane, T. J. and D. O. Staiger (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives 16*(4), 91–114.

Kane, T. J. and D. O. Staiger (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Kane, T. J. and D. O. Staiger (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Technical report, Bill & Melinda Gates Foundation MET Project Research Paper.

Kane, T. J., D. O. Staiger, D. Grissmer, and H. F. Ladd (2002). Volatility in school test scores: Implications for test-based accountability systems. *Brookings Papers on Education Policy* (5), 235–283.

Kane, T. J., E. S. Taylor, J. H. Tyler, and A. L. Wooten (2011). Identifying effective classroom practices using student achievement data. *Journal of human Resources 46*(3), 587–613.

Koedel, C., K. Mihaly, and J. E. Rockoff (2015). Value-added modeling: A review. *Economics of Education Review 47*, 180–195.

Koretz, D. M. et al. (1996). Perceived effects of the Kentucky Instructional Results Information System (KIRIS). Technical report, RAND Corporation Institute on Education and Training.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The quarterly journal of economics 114*(2), 497–532.

Krueger, A. B. (2003). Economic considerations and class size. *The Economic Journal 113*(485), F34–F63.

Macartney, H., R. McMillan, and U. Petronijevic (2018). Teacher value-added and economic agency. NBER Working Paper No. 24747.

Macartney, H., R. McMillan, and U. Petronijevic (2021). A quantitative framework for analyzing the distributional effects of incentive schemes. NBER Working Paper No. 28816.

Martin, A. J. and G. Lazendic (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology 110*(1), 27.

McGiverin, J., D. Gilman, and C. Tillitski (1989). A meta-analysis of the relation between class size and achievement. *The Elementary School Journal 90*(1), 47–56.

Molnar, A., P. Smith, J. Zahorik, A. Palmer, A. Halbach, and K. Ehrle (1999). Evaluating the sage program: A pilot program in targeted pupil-teacher reduction in wisconsin. *Educational Evaluation and Policy Analysis 21*(2), 165–177.

Neal, D. and D. W. Schanzenbach (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics 92*(2), 263–283.

Nielsen, E. (2019). Test questions, economic outcomes, and inequality. Working paper.

Pedulla, J. J., L. M. Abrams, G. F. Madaus, M. K. Russell, M. A. Ramos, and J. Miao (2003). Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers. Technical report, National Board on Educational Testing and Public Policy.

Pinker, S. (1997). *How the Mind Works*. WW Norton & Company.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics 92*(5-6), 1394–1415.

Reback, R., J. Rockoff, and H. L. Schwartz (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. *American Economic Journal: Economic Policy 6*(3), 207–41.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

Rockoff, J. and L. J. Turner (2010). Short-run impacts of accountability on school quality. *American Economic Journal: Economic Policy 2*(4), 119–47.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics 125*(1), 175–214.

Rouse, C. E., J. Hannaway, D. Goldhaber, and D. Figlio (2013). Feeling the Florida heat? how low-performing schools respond to voucher and accountability pressure. *American*

Economic Journal: Economic Policy 5(2), 251–81.

Schanzenbach, D. W. (2014). Does class size matter? Technical report, National Education Policy Center.

Springer, M. G. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review 27*(5), 556–563.

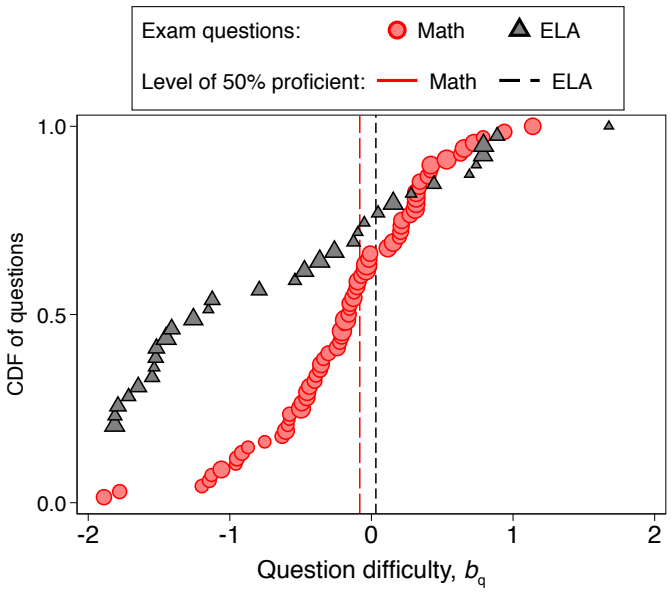
Stecher, B. M. and K. J. Mitchell (1995). Portfolio-driven reform: Vermont teachers' understanding of mathematical problem solving and related changes in classroom practice. Technical report, National Center for Research on Evaluation, Standards, and Student Testing.

Taylor, E. S. and J. H. Tyler (2012). The effect of evaluation on teacher performance. *American Economic Review 102*(7), 3628–51.

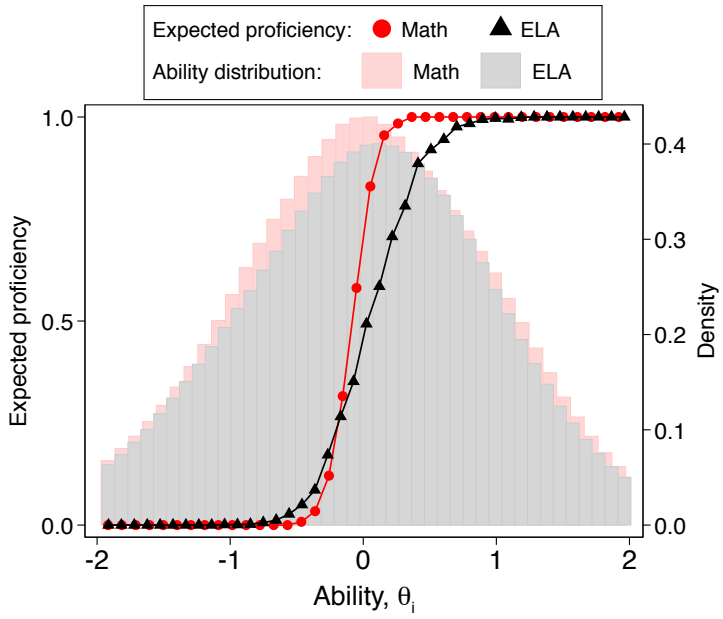
Topol, B., J. Olson, E. Roeber, and P. Hennon (2012). Getting to higher-quality assessments: Evaluating costs, benefits, and investment strategies. Technical report, Stanford Center for Opportunity Policy in Education.

Urquiola, M. (2006). Identifying class size effects in developing countries: Evidence from rural Bolivia. *Review of Economics and Statistics 88*(1), 171–177.

Wong, M., T. D. Cook, and P. M. Steiner (2009). No Child Left Behind: An interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series. Institute for Policy Research Working Paper Series.



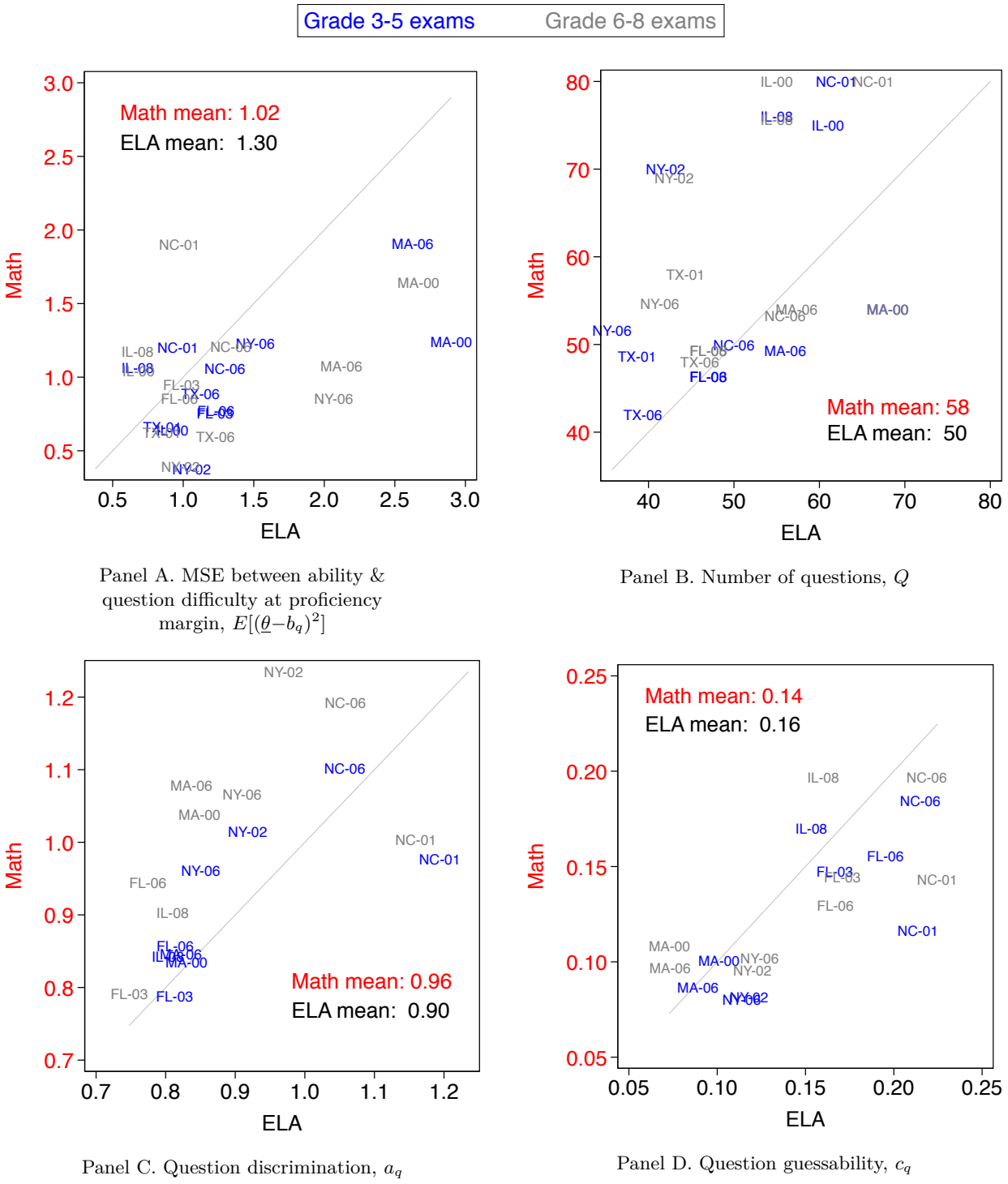
Panel A. Question difficulty



Panel B. Expected proficiency

FIGURE 1. New York 8th grade exams (2006)

Notes: This figure presents statistics on New York 8th grade math and ELA exams in 2006. In Panel A, markers depict the cumulative distribution function of question difficulty, b_q; we do not plot ELA questions with $b_q < -2$ to improve readability. Vertical dash lines represent the level of ability, $\underline{\theta}$, at which expected proficiency is 50 percent in our simulations. In Panel B, lines depict expected proficiency (left axis) at 0.1 increment of ability, θ_i, from our simulations. Vertical bars show the density of the ability distribution (right axis) at each ability level.



Panel A. MSE between ability & question difficulty at proficiency margin, $E[(\theta - b_q)^2]$

Panel B. Number of questions, Q

Panel C. Question discrimination, a_q

Panel D. Question guessability, c_q

FIGURE 2. Exam design parameters

Notes: This figure plots exam design parameters in math (y-axes) against parameters for the ELA exam (x-axes) in the same state, year, and grade group. The text of each marker indicates the state and year (e.g., MA-06), and the color indicates the grade group for which the statistics are averaged (blue = 3–5; grey = 6–8). The statistic for each graph is listed in the panel title; see Appendix Table A1 for details. In Panel A, we moved the value for the MA-2000 grade 3–5 ELA exam from 4.1 to 2.9 to improve readability.

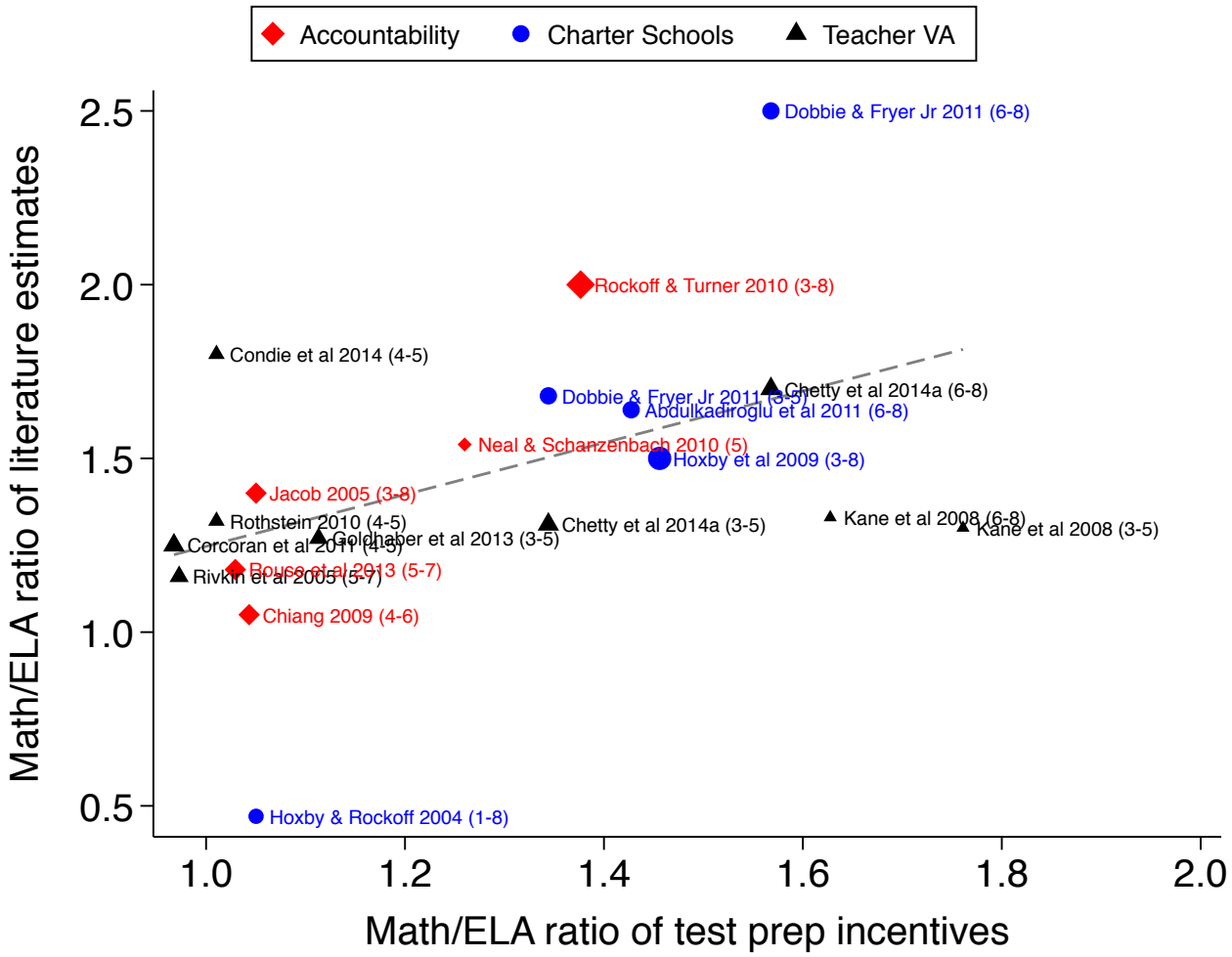
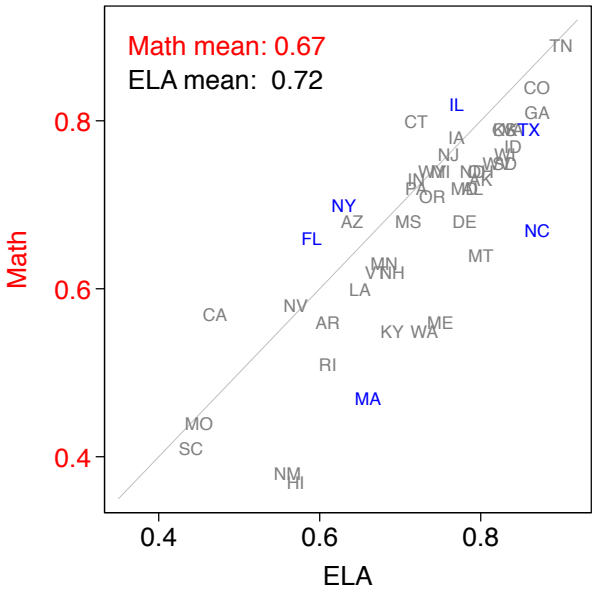
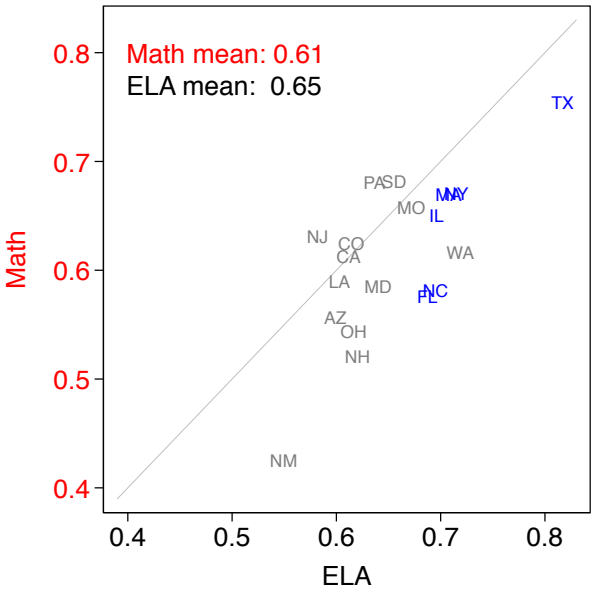


FIGURE 3. Math/ELA ratios of test prep incentives and literature estimates

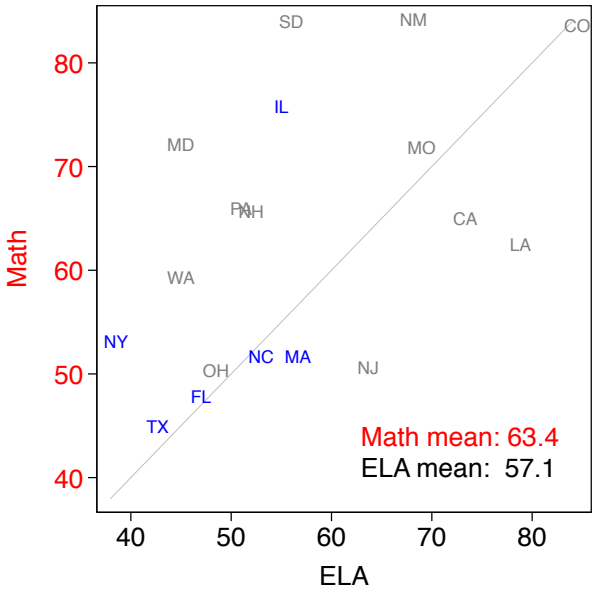
Notes: This figure plots the math/ELA ratios of literature estimates and test prep incentives for the exams in our sample. The y-axis is the math/ELA ratio of literature estimates, as in column (H) of Table 1. We include the 19 literature estimates for which the state matches that in our sample of exams. The x-axis is the math/ELA ratio of the density of the ability distribution $\times$ the derivative of expected proficiency (computed at the proficiency margin), as in column (G) of Table 3. This value averages over the grades (3–8) and periods (pre-NCLB and NCLB era) that match the sample for each literature estimate. We report the grade range for the matched samples in parentheses. The dashed line shows the OLS relationship between the two variables. Marker sizes are proportional to the number of exams in our sample that match the authors' data range. We moved the value of the literature estimate for grades 6–8 in Dobbie and Fryer Jr (2011) from 4.87 to 2.5 to improve readability.



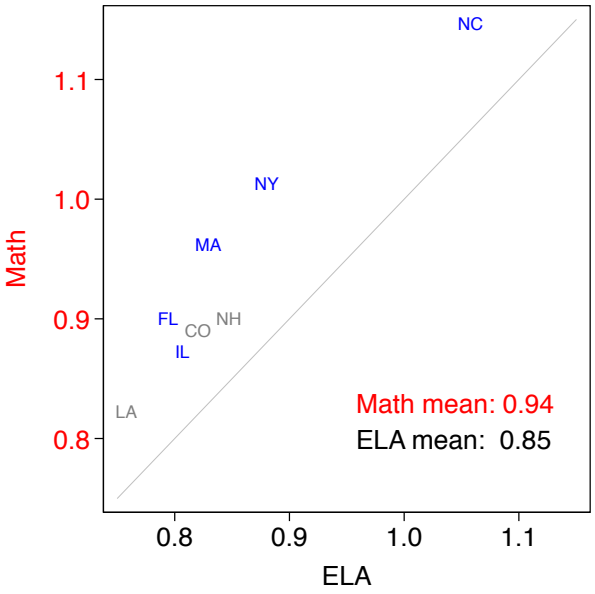
Panel A. Proficiency rate



Panel B. Proportion correct



Panel C. Number of questions, Q



Panel D. Question discrimination, a_q

FIGURE 4. Statistics and design parameters for NCLB era state exams (2006–2009)

Notes: This figure plots exam statistics in math (y-axes) against statistics for the ELA exam (x-axes) in the same state. Panel A plots state proficiency rates for grade 4 and 8 exams in 2007 as reported by NAEP (available in October 2021 at: https://nces.ed.gov/nationsreportcard/studies/statemapping/2007_naep_state_table.asp). Panels B–D display the proportion of correct answers, number of questions, and mean question discrimination from technical reports we could find for states in the NCLB era (see Appendix Table C1). Statistics are averaged across all grade levels.

TABLE 1. Literature estimates

(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
					Main effects		Ratio
Paper	State	District (if not state)	Years	Grades	Math	ELA	(F)/(G)
Panel A. Accountability effects (Figlio and Loeb, 2011)							
Chiang (2009)	FL		2002–03	4–6	0.12	0.11	1.05
Rouse et al. (2013)	FL		2003–05	5–7	0.09	0.08	1.18
Jacob (2005)	IL	Chicago	1993–00	3–8	0.33	0.24	1.40
Neal and Schanzenbach (2010)	IL	Chicago	2001–02	5	0.06	0.04	1.54
Rockoff and Turner (2010)	NY	New York City	2007–08	3–8	0.10	0.05	2.00
Wong et al. (2009)	National		1990–09	4	0.41	0.19	2.16
Dee and Jacob (2011)	National		1992–07	4	0.23	0.06	3.66
Average					**0.19**	**0.11**	**1.75**
Panel B. Charter school effects (Epple et al., 2016)							
Curto and Fryer Jr (2014)	DC		2008–09	7–8	0.22	0.20	1.08
Hoxby and Rockoff (2004)	IL	Chicago	2000–02	1–8	0.07	0.15	0.47
Abdulkadiroğlu et al. (2011)	MA	Boston	2002–08	6–8	0.42	0.25	1.64
Hoxby et al. (2009)	NY	New York City	2001–08	3–8	0.09	0.06	1.50
Dobbie and Fryer Jr (2011)	NY	New York City	2004–10	3–5	0.19	0.11	1.68
Dobbie and Fryer Jr (2011)	NY	New York City	2004–10	6–8	0.23	0.05	4.87
Gleason et al. (2010)	National		2005–08	4–8	-0.08	-0.10	0.80
Average					**0.16**	**0.10**	**1.57**
Panel C. Standard deviation of teacher value added (Koedel et al., 2015)							
Kane and Staiger (2008)	CA	Los Angeles	2000–03	2–5	0.22	0.18	1.25
Bacher-Hicks et al. (2014)	CA	Los Angeles	2005–11	3–5	0.29	0.19	1.52
Bacher-Hicks et al. (2014)	CA	Los Angeles	2005–11	6–8	0.21	0.10	2.12
Rothstein (2010)	NC		2000–01	4–5	0.15	0.11	1.32
Condie et al. (2014)	NC		1998–04	4–5	0.18	0.10	1.80
Goldhaber et al. (2013)	NC		1999–05	3–5	0.24	0.19	1.27
Kane et al. (2008)	NY	New York City	1999–05	3–5	0.13	0.10	1.30
Kane et al. (2008)	NY	New York City	1999–05	6–8	0.08	0.06	1.33
Rivkin et al. (2005)	TX		1995–98	5–7	0.11	0.10	1.16
Corcoran et al. (2011)	TX	Houston	1999–06	4–5	0.26	0.21	1.25
Jacob and Lefgren (2008)		"Western U.S. district"	1998–05	2–6	0.26	0.12	2.17
Chetty et al. (2014a)		"Large urban district"	1989–09	3–5	0.16	0.12	1.31
Chetty et al. (2014a)		"Large urban district"	1989–09	6–8	0.13	0.08	1.70
Average					**0.19**	**0.13**	**1.47**

Notes: This table shows math and ELA estimates from studies cited in three literature review papers. Panel A includes papers on the effects of accountability policies cited in Figlio and Loeb (2011) (their Tables 8.1–8.2). Panel B includes papers on the effects of charter school admission from Epple et al. (2016) (their Table 6). Panel C includes estimates of the standard deviation of teacher value added from papers cited throughout Koedel et al. (2015).

We include only papers that present effects on grade 3–8 scores in *both* math and ELA. We also exclude papers with identification strategies that are typically perceived as less credible by researchers in the economics of education; these include empirical designs based on control for observables, student fixed effects, and propensity score matching.

Column (A) lists the papers that meet these criteria. Columns (B)–(E) show the state, school district (if not statewide), exam years (spring of academic year), and grades for the sample(s) in each paper. Columns (F)–(G) show math and ELA estimates from the authors' benchmark specification. Most estimates are in units of standardized test scores (mean 0/SD 1), but the units differ in a few papers. In all cases, the math and ELA effects are from the same specification and in the same units. Column (H) shows the ratio of columns (F) and (G).

TABLE 2. Summary statistics for all exam takers

			(A)	(B)	(C)	(D)	(E)	(F)
			Proficiency rate		Proportion correct answers		Derivative of proportion correct	
State	Grades	Year	Math	ELA	Math	ELA	Math	ELA
Panel A. All exams ($N = 122$)								
All exam mean			**0.66**	**0.70**	**0.63**	**0.70**	**0.18**	**0.16**
Panel B. Pre-NCLB exams ($N = 50$)								
Florida	3-5	2003	0.56	0.60	0.57	0.64	0.18	0.16
Illinois	3,5	2000	0.59	0.54	0.58	0.62	0.18	0.17
Massachusetts	4	2000	0.40	0.20	0.57	0.61	0.19	0.15
New York	4	2002	0.68	0.64	0.66	0.70	0.25	0.19
North Carolina	3-5	2001	0.80	0.74	0.62	0.68	0.21	0.19
Texas	3-5	2001	0.89	0.89	0.76	0.83	0.14	0.11
Grade 3-5 mean			**0.69**	**0.66**	**0.63**	**0.69**	**0.18**	**0.16**
Florida	6-8	2003	0.50	0.51	0.49	0.64	0.16	0.16
Illinois	8	2000	0.42	0.61	0.55	0.64	0.18	0.18
Massachusetts	8	2000	0.34	0.61	0.50	0.61	0.19	0.15
New York	8	2002	0.38	0.47	0.54	0.61	0.29	0.21
North Carolina	6-8	2001	0.79	0.73	0.52	0.68	0.20	0.19
Texas	6-8	2001	0.91	0.89	0.75	0.82	0.14	0.12
Grade 6-8 mean			**0.65**	**0.67**	**0.57**	**0.69**	**0.18**	**0.16**
Panel C. NCLB era exams ($N = 72$)								
Florida	3-5	2006	0.65	0.70	0.60	0.66	0.18	0.15
Illinois	3-5	2008	0.86	0.75	0.68	0.67	0.18	0.18
Massachusetts	3-5	2006	0.45	0.56	0.71	0.71	0.17	0.14
New York	3-5	2006	0.76	0.68	0.72	0.72	0.17	0.15
North Carolina	3-5	2006	0.63	0.84	0.59	0.67	0.20	0.18
Texas	3-5	2006	0.82	0.83	0.80	0.81	0.13	0.12
Grade 3-5 mean			**0.70**	**0.73**	**0.68**	**0.71**	**0.17**	**0.16**
Florida	6-8	2006	0.56	0.57	0.52	0.65	0.19	0.16
Illinois	6-8	2008	0.82	0.80	0.63	0.70	0.19	0.18
Massachusetts	6-8	2006	0.42	0.68	0.63	0.70	0.21	0.15
New York	6-8	2006	0.57	0.56	0.60	0.70	0.22	0.15
North Carolina	6-8	2006	0.60	0.82	0.54	0.66	0.19	0.19
Texas	6-8	2006	0.72	0.84	0.70	0.82	0.16	0.12
Grade 6-8 mean			**0.61**	**0.71**	**0.61**	**0.70**	**0.19**	**0.16**

Notes: This table shows summary statistics for all exam takers from our simulations. Panel A reports averages across all 122 exams in our sample. Panels B–C report averages by state, exam year (pre-NCLB and NCLB era), and grade group (3–5 and 6–8). Columns (A)–(B) show the proportion of exam takers who achieved proficiency, computed separately for math and ELA exams. Columns (C)–(D) show the mean fraction of questions that exam takers answered correctly. Columns (E)–(F) show the derivative of the probability of a correct answer, $p'_q(\theta_i)$, averaged across all questions and exam takers.

TABLE 3. Test prep incentives for exam takers on proficiency margin ($\theta_i = \underline{\theta}$)

			(A)	(B)	(C)	(D)	(E)	(F)	(G)
			Density of ability distribution		Derivative of expected proficiency		Density × derivative		Ratio
State	Grades	Year	Math	ELA	Math	ELA	Math	ELA	(E)/(F)
Panel A. All exams ($N = 122$)									
All exam mean			**0.32**	**0.30**	**1.50**	**1.32**	**0.49**	**0.40**	**1.22**
Panel B. Pre-NCLB exams ($N = 50$)									
Florida	3-5	2003	0.33	0.32	1.36	1.35	0.46	0.44	1.05
Illinois	3,5	2000	0.40	0.37	1.55	1.46	0.62	0.54	1.15
Massachusetts	4	2000	0.38	0.27	1.51	1.35	0.57	0.37	1.55
New York	4	2002	0.45	0.42	2.42	1.47	1.09	0.62	1.76
North Carolina	3-5	2001	0.35	0.34	1.61	1.50	0.57	0.51	1.11
Texas	3-5	2001	0.15	0.14	1.31	1.18	0.19	0.16	1.18
Grade 3-5 mean			**0.32**	**0.30**	**1.53**	**1.37**	**0.50**	**0.42**	**1.21**
Florida	6-8	2003	0.34	0.33	1.23	1.31	0.41	0.43	0.96
Illinois	8	2000	0.35	0.45	1.51	1.38	0.53	0.62	0.86
Massachusetts	8	2000	0.36	0.36	1.73	1.47	0.63	0.53	1.17
New York	8	2002	0.48	0.48	2.67	1.64	1.29	0.79	1.63
North Carolina	6-8	2001	0.34	0.34	1.38	1.50	0.48	0.51	0.93
Texas	6-8	2001	0.14	0.14	1.47	1.22	0.21	0.17	1.22
Grade 6-8 mean			**0.31**	**0.31**	**1.51**	**1.38**	**0.48**	**0.44**	**1.09**
Panel C. NCLB era exams ($N = 72$)									
Florida	3-5	2006	0.31	0.27	1.40	1.25	0.44	0.34	1.28
Illinois	3-5	2008	0.22	0.33	1.48	1.39	0.32	0.46	0.68
Massachusetts	3-5	2006	0.39	0.38	1.40	1.38	0.55	0.53	1.04
New York	3-5	2006	0.27	0.33	1.66	1.14	0.46	0.38	1.21
North Carolina	3-5	2006	0.39	0.25	1.33	1.20	0.53	0.30	1.77
Texas	3-5	2006	0.22	0.21	1.21	1.11	0.27	0.23	1.14
Grade 3-5 mean			**0.30**	**0.30**	**1.41**	**1.25**	**0.43**	**0.37**	**1.14**
Florida	6-8	2006	0.34	0.33	1.53	1.27	0.53	0.41	1.27
Illinois	6-8	2008	0.25	0.29	1.46	1.36	0.36	0.40	0.92
Massachusetts	6-8	2006	0.38	0.33	1.84	1.42	0.69	0.47	1.48
New York	6-8	2006	0.39	0.38	1.91	1.24	0.75	0.48	1.58
North Carolina	6-8	2006	0.40	0.28	1.34	1.30	0.53	0.37	1.44
Texas	6-8	2006	0.31	0.19	1.30	1.20	0.41	0.23	1.74
Grade 6-8 mean			**0.34**	**0.30**	**1.56**	**1.30**	**0.55**	**0.39**	**1.39**

Notes: This table shows statistics on test prep incentives for exam takers on proficiency margin ($\theta_i = \underline{\theta}$). Panel A reports averages across all 122 exams in our sample. Panels B–C report averages by state, exam year (pre-NCLB and NCLB era), and grade group (3–5 and 6–8). We define the proficiency margin, $\underline{\theta}$, as the level of ability where the likelihood of proficiency is 50 percent. Columns (A)–(B) show the density of the ability distribution at $\underline{\theta}$, computed separately for math and ELA exams. Columns (C)–(D) show the derivative of expected proficiency $dE[\tau(R_i)|\theta_i]/d\theta_i$ at $\underline{\theta}$. Column (E) shows the product of the terms in columns (A) and (C). Column (F) shows the product of the terms in columns (B) and (D). Column (G) shows the ratio of columns (E) and (F).

TABLE 4. Math/ELA differences in test prep incentives and exam design

Dependent variable	(A) ELA mean	(B) All exams	(C) Grades 3–5	(D) Grades 6–8	(E) Pre-NCLB	(F) NCLB era
		Math − ELA difference				
Panel A. Statistics for all exam takers						
Proficiency rate	0.70	−0.04**	−0.00	−0.07***	0.00	−0.06***
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Proportion correct	0.70	−0.07***	−0.04***	−0.11***	−0.09***	−0.06***
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Derivative of proportion correct	0.16	0.02***	0.02***	0.03***	0.02***	0.03***
		(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Panel B. Statistics for exam takers on proficiency margin ($\theta_i = \underline{\theta}$)						
Density of ability distribution	0.30	0.02**	0.01	0.02**	0.01	0.02*
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Derivative of expected proficiency	1.32	0.19***	0.16***	0.21***	0.14**	0.22***
		(0.04)	(0.04)	(0.06)	(0.06)	(0.04)
Density of ability distribution × derivative of expected proficiency	0.40	0.09***	0.07**	0.11***	0.06**	0.10***
		(0.02)	(0.03)	(0.03)	(0.03)	(0.02)
Panel C. Exam design parameters						
Number of questions, Q	50.35	7.28***	7.12***	7.44***	10.00***	5.39***
		(1.50)	(2.26)	(1.99)	(2.43)	(1.85)
Question discrimination, a_q	0.90	0.06***	0.01	0.11***	−0.03	0.11***
		(0.02)	(0.02)	(0.03)	(0.04)	(0.01)
Question difficulty, b_q	−0.67	0.34***	0.14**	0.53***	0.33***	0.34***
		(0.04)	(0.05)	(0.04)	(0.07)	(0.05)
Question guessability, c_q	0.16	−0.02***	−0.03***	−0.02*	−0.04***	−0.01
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
MSE between ability/question difficulty at prof. margin, $E[(\underline{\theta}-b_q)^2]$	1.30	−0.28***	−0.32**	−0.24	−0.19	−0.35***
		(0.09)	(0.12)	(0.15)	(0.15)	(0.12)
N (# exams)	61	122	62	60	50	72

Notes: This table shows results from a regression of exam characteristics on an indicator for math exams (relative to ELA exams). The dependent variable for each regression is listed in the first column; these variables are defined as in Tables 2–3 and Figure 2. Column (A) shows the mean of each dependent variable for the 61 ELA exams in our sample. Column (B) displays the coefficient on an indicator for math exams in regressions that include all 122 exams in our sample. Columns (C)–(F) show coefficients on the math indicator in subsamples that include only: grade 3–5 exams, grade 6–8 exams, pre-NCLB exams (2000–2003), and NCLB era exams (2006–2008).

All regressions include fixed effects for state × year × grade triplets. Parentheses contain robust standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 5. Predictors of the derivative of expected proficiency for marginally-proficient students

	Dependent variable: Derivative of expected proficiency for marginally-proficient students, i.e., $dE[\tau(R_i)\|\theta_i]/d\theta_i$ at $\theta_i = \underline{\theta}$					
Covariate (standardized to mean 0/SD 1)	(A)	(B)	(C)	(D)	(E)	(F)
Constant	1.41^{***}	1.41^{***}	1.44^{***}	1.44^{***}	1.41^{***}	1.45^{***}
	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)
MSE between ability/question difficulty at prof. margin, $E[(\underline{\theta}-b_q)^2]$	-0.05^{*}				-0.07^{***}	-0.12^{***}
	(0.02)				(0.02)	(0.02)
Number of questions, Q		0.13^{***}			0.14^{***}	0.11^{***}
		(0.02)			(0.02)	(0.02)
Question discrimination, a_q			0.12^{***}			0.12^{***}
			(0.04)			(0.02)
Question guessability, c_q				-0.11^{***}		-0.18^{***}
				(0.03)		(0.02)
N (# exams)	122	122	92	92	122	92

Notes: This table shows results from regressions in which the dependent variable is the derivative of expected proficiency for marginally-proficient students (defined as in columns C–D of Table 3). The covariates for each regression are listed in the first column; these are the exam design features in Figure 2. We standardize each covariate to be mean zero and SD one. The SDs for each characteristic are: MSE between ability/question difficulty at prof. margin (0.63); number of questions (12); question discrimination (0.14); question guessability (0.05).

The sample for each regression includes the 122 exams in our main sample. Question discrimination (a_q) and guessability (c_q) are not defined for exams that use the Rasch model (see Appendix Table A1). Parentheses contain robust standard errors.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 6. Math/ELA differences in test prep incentives by bandwidth around proficiency margin

	(A)	(B)	(C)	(D)	(E)	(F)
		Math − ELA difference by bandwidth around $\underline{\theta}$				
Dependent variable	ELA mean at $\theta_i = \underline{\theta}$	$\theta_i = \underline{\theta}$	$\lvert\theta_i - \underline{\theta}\rvert < 0.2$	$\lvert\theta_i - \underline{\theta}\rvert < 0.4$	$\lvert\theta_i - \underline{\theta}\rvert < 0.6$	$\lvert\theta_i - \underline{\theta}\rvert < 0.8$
Density of ability distribution	0.30	0.02^{**}	0.02^{**}	0.02^{**}	0.02^{*}	0.01^{*}
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Derivative of expected proficiency	1.32	0.19^{***}	0.16^{***}	0.07^{***}	0.02^{***}	-0.00
		(0.04)	(0.03)	(0.01)	(0.01)	(0.00)
Density of ability distribution × derivative of expected proficiency	0.40	0.09^{***}	0.08^{***}	0.04^{***}	0.02^{**}	0.01
		(0.02)	(0.02)	(0.01)	(0.01)	(0.00)
N (# exams)	61	122	122	122	122	122
Expected proficiency at minimum θ_i	0.50	0.50	0.30	0.12	0.04	0.01
Expected proficiency at maximum θ_i	0.50	0.50	0.70	0.89	0.97	0.99

Notes: This table shows results from a regression of exam characteristics on an indicator for math exams (relative to ELA exams). The dependent variable for each regression is listed in the first column; these variables are defined as in Panel B of Table 4. Columns (A)–(B) show values for exam takers on the proficiency margin, i.e., those with ability $\theta_i = \underline{\theta}$. Column (A) shows the mean of each dependent variable in the 61 ELA exams in our sample. Column (B) displays the coefficient on an indicator for math exams in regressions that include all 122 exams in our sample. Columns (C)–(F) are analogous to column (B), but the values represent averages over test takers whose ability, θ_i, is within a certain bandwidth around the proficiency margin, $\underline{\theta}$. These bandwidths are 0.2, 0.4, 0.6, and 0.8 units on the θ_i scale, as listed in the column header. The bottom two rows show expected proficiency at the minimum and maximum values of θ_i within each bandwidth.

All regressions include fixed effects for state × year × grade triplets. Parentheses contain robust standard errors.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 7. Regressions of math/ELA ratio of literature estimates on math/ELA ratio of test prep incentives

	(A)	(B)	(C)	(D)
	Dependent variable: Math/ELA ratio of literature estimates			
Covariate	All papers	Account-ability	Charter school	Teacher VA
Panel A. Weighted by number of exams				
Math/ELA ratio of test prep incentives	0.83***	2.33***	1.51***	0.32
	(0.23)	(0.27)	(0.31)	(0.27)
N (# exams)	63	16	23	24
Panel B. Weighted by number of literature estimates				
Math/ELA ratio of test prep incentives	0.51*	2.16***	1.61***	0.08
	(0.28)	(0.44)	(0.24)	(0.24)
N (# literature estimates)	19	5	5	9
Dependent variable mean	1.45	1.43	1.57	1.38

Notes: This table shows the OLS relationship between the math/ELA ratios of literature estimates and test prep incentives. The dependent variable for each regression is the math/ELA ratio of literature estimates, as in column (H) of Table 1. We include the 19 literature estimates for which the authors' data comes from a state in our sample of exams. To reduce noise in the regressions, we winsorize the dependent variable by setting the maximum/minimum literature estimates to the 90th/10th percentile values (i.e., 4.87 becomes 2, and 0.47 becomes 1.05). The only covariate in each regression is the math/ELA ratio of the density of the ability distribution $\times$ the derivative of expected proficiency (computed at the proficiency margin), as in column (G) of Table 3. This value averages over the grades (3–8) and periods (pre-NCLB and NCLB era) that match the sample for each literature estimate.

Regressions in column (A) include all 19 matched literature estimates. Columns (B)–(D) estimate regressions separately for the accountability, charter school, and teacher value added papers in this matched sample (see Table 1). Panel A presents results from regressions weighted by the number of matched exams. Panel B presents results from regressions weighted by the number of literature estimates. Parentheses in both panels contain standard errors clustered at the literature estimate level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix — For Online Publication

Outline:

A. Appendix figures and tables
B. Theoretical appendix
C. Empirical appendix

A. Appendix figures and tables

Table A1. Exam design parameters

		(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)
			MSE between ability/difficulty on prof. margin $E[(\underline{\theta}-b_q)^2]$		Number of questions Q		Question discrimination a_q		Question guessability c_q	
State & exam grades		Rasch?	Math	ELA	Math	ELA	Math	ELA	Math	ELA
Panel A. All exams ($N = 122$)										
All exam mean			**1.02**	**1.30**	**57.6**	**50.3**	**0.96**	**0.90**	**0.14**	**0.16**
Panel B. Pre-NCLB exams (2000–2003, $N = 50$)										
Florida	3-5		0.76	1.23	46.3	47.0	0.79	0.81	0.15	0.17
Illinois	3,5	✓	0.64	0.92	75.0	61.0	.	.	.	.
Massachusetts	4		1.24	4.12	54.0	68.0	0.84	0.83	0.10	0.10
New York	4		0.38	1.06	70.0	42.0	1.02	0.92	0.08	0.12
North Carolina	3-5		1.20	0.96	80.0	62.0	0.98	1.19	0.12	0.21
Texas	3-5	✓	0.66	0.85	48.7	38.7	.	.	.	.
Grade 3-5 mean			**0.83**	**1.24**	**61.5**	**51.9**	**0.89**	**0.97**	**0.12**	**0.17**
Florida	6-8		0.95	0.99	49.3	47.0	0.79	0.75	0.14	0.17
Illinois	8	✓	1.05	0.68	80.0	55.0	.	.	.	.
Massachusetts	8		1.64	2.67	54.0	68.0	1.04	0.85	0.11	0.07
New York	8		0.39	0.98	69.0	43.0	1.23	0.97	0.10	0.12
North Carolina	6-8		1.90	0.97	80.0	66.3	1.00	1.16	0.14	0.22
Texas	6-8	✓	0.62	0.85	58.0	44.3	.	.	.	.
Grade 6-8 mean			**1.13**	**1.06**	**63.8**	**53.3**	**0.96**	**0.94**	**0.13**	**0.17**
Panel C. NCLB era exams (2006–2008, $N = 72$)										
Florida	3-5		0.77	1.23	46.3	47.0	0.86	0.81	0.16	0.20
Illinois	3-5		1.07	0.68	76.0	55.0	0.84	0.80	0.17	0.15
Massachusetts	3-5		1.91	2.63	49.3	56.0	0.85	0.82	0.09	0.09
New York	3-5		1.23	1.51	51.7	35.7	0.96	0.85	0.08	0.11
North Carolina	3-5		1.06	1.30	50.0	50.0	1.10	1.06	0.18	0.21
Texas	3-5	✓	0.89	1.12	42.0	39.3	.	.	.	.
Grade 3-5 mean			**1.15**	**1.41**	**52.6**	**47.2**	**0.92**	**0.87**	**0.14**	**0.15**
Florida	6-8		0.86	0.97	49.3	47.0	0.94	0.77	0.13	0.17
Illinois	6-8		1.17	0.68	75.7	55.0	0.90	0.81	0.20	0.16
Massachusetts	6-8		1.08	2.12	54.0	57.3	1.08	0.84	0.10	0.07
New York	6-8		0.86	2.07	54.7	41.3	1.07	0.91	0.10	0.12
North Carolina	6-8		1.21	1.34	53.3	56.0	1.19	1.06	0.20	0.22
Texas	6-8	✓	0.60	1.23	48.0	46.0	.	.	.	.
Grade 6-8 mean			**0.96**	**1.40**	**55.8**	**50.4**	**1.04**	**0.88**	**0.14**	**0.15**

Notes: This table shows the values of the exam design parameters that are plotted in Figure 2. The table has the same structure as Tables 2–3. A checkmark in column (A) indicates that the exams were scored using the Rasch model, which does not estimate question discrimination (a_q) or guessability (c_q). See the notes to Figure 2 for details.

TABLE A2. Predictors of the derivative of expected proficiency for marginally-proficient students with square terms

	Dependent variable: Derivative of expected proficiency for marginally-proficient students, i.e., $dE[\tau(R_i)\|\theta_i]/d\theta_i$ at $\theta_i = \underline{\theta}$					
Covariate (standardized to mean 0/SD 1)	(A)	(B)	(C)	(D)	(E)	(F)
Constant	1.38***	1.46***	1.44***	1.41***	1.44***	1.44***
	(0.02)	(0.03)	(0.04)	(0.03)	(0.02)	(0.04)
MSE between ability/question difficulty at prof. margin, $E[(\underline{\theta}-b_q)^2]$	−0.11**				−0.10**	−0.18***
	(0.05)				(0.04)	(0.03)
MSE^2	0.03*				0.02	0.02***
	(0.01)				(0.01)	(0.01)
Number of questions, Q		0.17***			0.17***	0.12***
		(0.02)			(0.02)	(0.02)
Number of questions^2		−0.05***			−0.05***	−0.03*
		(0.01)			(0.01)	(0.02)
Question discrimination, a_q			0.12***			0.11***
			(0.03)			(0.02)
Question discrimination^2			0.01			0.02
			(0.04)			(0.02)
Question guessability, c_q				−0.11***		−0.18***
				(0.03)		(0.02)
Question guessability^2				0.04		0.01
				(0.03)		(0.02)
N (# exams)	122	122	92	92	122	92

Notes: This table shows results from regressions in which the dependent variable is the derivative of expected proficiency for marginally-proficient students (defined as in columns C–D of Table 3). This table is similar to Table 5, except we include square terms for each exam design parameter.

The covariates for each regression are listed in the first column; these are the exam design features in Figure 2 and their squares. We standardize each covariate to be mean zero and SD one. The SDs for each characteristic are: MSE between ability/question difficulty at prof. margin (0.63); number of questions (12); question discrimination (0.14); question guessability (0.05).

The sample for each regression includes the 122 exams in our main sample. Question discrimination (a_q) and guessability (c_q) are not defined for exams that use the Rasch model (see Appendix Table A1). Parentheses contain robust standard errors.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A3. Robustness checks for math/ELA differences in test prep incentives and exam design

	(A)	(B)	(C)	(D)	(E)
	Math − ELA difference				
Dependent variable	Full sample	IRT param.	p-values	Raw to scale	$\theta_i \sim N(0,1)$
Panel A. Statistics for all exam takers					
Proficiency rate	−0.04**	−0.05*	−0.02	−0.03	−0.02
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)
Proportion correct	−0.07***	−0.06***	−0.07***	−0.05***	−0.06***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Derivative of proportion correct	0.02***	0.03***	0.03***	0.03***	0.02***
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)
Panel B. Statistics for exam takers on proficiency margin ($\theta_i = \underline{\theta}$)					
Density of ability distribution	0.02**	0.01	0.03*	0.00	0.00
	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)
Derivative of expected proficiency	0.19***	0.30***	0.26***	0.28***	0.19***
	(0.04)	(0.06)	(0.09)	(0.05)	(0.04)
Density of ability distribution × derivative of expected proficiency	0.09***	0.12***	0.13**	0.10***	0.06***
	(0.02)	(0.03)	(0.05)	(0.03)	(0.02)
Panel C. Exam design parameters					
Number of questions, Q	7.28***	3.82	9.93***	8.64***	7.28***
	(1.50)	(2.80)	(2.35)	(2.11)	(1.50)
Question discrimination, a_q	0.06***	0.12***	0.18	0.12***	0.06***
	(0.02)	(0.02)	(0.08)	(0.02)	(0.02)
Question difficulty, b_q	0.34***	0.31***	0.44***	0.34***	0.34***
	(0.04)	(0.07)	(0.07)	(0.06)	(0.04)
Question guessability, c_q	−0.02***	−0.01*	−0.03	0.00	−0.02***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
MSE between ability/question difficulty at prof. margin, $E[(\underline{\theta}-b_q)^2]$	−0.28***	−0.68***	−0.37**	−0.47***	−0.28***
	(0.09)	(0.17)	(0.12)	(0.14)	(0.09)
N (# exams)	122	46	28	68	122

Notes: This table presents robustness checks for our main results. The dependent variable for each regression is listed in the first column; these variables are defined as in Tables 2–3 and Figure 2. Column (A) displays the coefficient on an indicator for math exams in regressions that include all 122 exams in our sample; this replicates our benchmark results from Table 4. Columns (B)–(D) show coefficients on the math indicator in subsamples of exams for which different types of data were available in the technical reports, as shown in Appendix Table C2. Column (B) includes exams where we observe question-level IRT data. Column (C) includes exams where we observe question-level p-values. Column (D) includes exams where we observe raw-to-scale conversion data. Column (E) displays the math coefficients from regressions with all exams, but we weight test taker ability, θ_i, in our simulated data by the standard normal distribution (rather than using weights that match the realized distribution of test scores). This shows how our results would change if the exams were offered to the reference population rather than the population of test takers in the actual year of administration.

All regressions include fixed effects for state × year × grade triplets. Parentheses contain robust standard errors.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A4. Regressions of math/ELA ratio of literature estimates on math/ELA ratio of test prep incentives without winsorizing

	(A)	(B)	(C)	(D)
	Dependent variable: Math/ELA ratio of literature estimates			
Covariate	All papers	Account-ability	Charter school	Teacher VA
Panel A. Weighted by number of exams				
Math/ELA ratio of test prep incentives	2.03*	2.33***	6.42	0.32
	(1.12)	(0.27)	(3.22)	(0.27)
N (# exams)	63	16	23	24
Panel B. Weighted by number of literature estimates				
Math/ELA ratio of test prep incentives	1.36	2.16***	6.56	0.08
	(0.96)	(0.44)	(3.08)	(0.24)
N (# literature estimates)	19	5	5	9
Dependent variable mean	1.57	1.43	2.03	1.38

Notes: This table shows the OLS relationship between the math/ELA ratios of literature estimates and test prep incentives. It is identical to Table 7, except we do not winsorize the dependent variable.

The dependent variable for each regression is the math/ELA ratio of literature estimates, as in column (H) of Table 1. We include the 19 literature estimates for which the authors' data comes from a state in our sample of exams. The only covariate in each regression is the math/ELA ratio of the density of the ability distribution × the derivative of expected proficiency (computed at the proficiency margin), as in column (G) of Table 3. This value averages over the grades (3–8) and periods (pre-NCLB and NCLB era) that match the sample for each literature estimate.

Regressions in column (A) include all 19 matched literature estimates. Columns (B)–(D) estimate regressions separately for the accountability, charter school, and teacher value added papers in this matched sample (see Table 1). Panel A presents results from regressions weighted by the number of matched exams. Panel B presents results from regressions weighted by the number of literature estimates. Parentheses in both panels contain standard errors clustered at the literature estimate level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A5. Regressions of math/ELA ratio of literature estimates on math/ELA ratio of test prep incentives dropping outliers

	(A)	(B)	(C)	(D)
	Dependent variable: Math/ELA ratio of literature estimates			
Covariate	All papers	Account-ability	Charter school	Teacher VA
Panel A. Weighted by number of exams				
Math/ELA ratio of test prep incentives	0.65**	2.33***	−1.54**	0.32
	(0.23)	(0.27)	(0.35)	(0.27)
N (# exams)	56	16	16	24
Panel B. Weighted by number of literature estimates				
Math/ELA ratio of test prep incentives	0.33	2.16***	−1.36	0.08
	(0.25)	(0.44)	(0.64)	(0.24)
N (# literature estimates)	17	5	3	9
Dependent variable mean	1.44	1.43	1.61	1.38

Notes: This table shows the OLS relationship between the math/ELA ratios of literature estimates and test prep incentives. It is identical to Table 7, except we drop the smallest and largest math/ELA ratios of literature estimates: Hoxby and Rockoff (2004) and Dobbie and Fryer Jr (2011) (grades 6–8).

The dependent variable for each regression is the math/ELA ratio of literature estimates, as in column (H) of Table 1. We include the 17 literature estimates for which the authors' data comes from a state in our sample of exams, excluding Hoxby and Rockoff (2004) and Dobbie and Fryer Jr (2011) (grades 6–8). The only covariate in each regression is the math/ELA ratio of the density of the ability distribution × the derivative of expected proficiency (computed at the proficiency margin), as in column (G) of Table 3. This value averages over the grades (3–8) and periods (pre-NCLB and NCLB era) that match the sample for each literature estimate.

Regressions in column (A) include all 17 matched literature estimates. Columns (B)–(D) estimate regressions separately for the accountability, charter school, and teacher value added papers in this matched sample (see Table 1). Panel A presents results from regressions weighted by the number of matched exams. Panel B presents results from regressions weighted by the number of literature estimates. Parentheses in both panels contain standard errors clustered at the literature estimate level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A6. Sensitivity of main results to simulation error

Dependent variable/covariate	(A) Main estimate	(B) Largest estimate	(C) Smallest estimate	(D) Prop. w/ $p < 0.05$	(E) Prop. w/ $p < 0.10$
		Statistics from 10 replications			
Panel A. Math/ELA differences in statistics for all exam takers					
Proficiency rate	-0.036^{**}	-0.037^{**}	-0.037^{**}	1.000	1.000
	(0.015)	(0.016)	(0.015)		
Proportion correct	-0.072^{***}	-0.072^{***}	-0.072^{***}	1.000	1.000
	(0.007)	(0.007)	(0.007)		
Derivative of proportion correct	0.024^{***}	0.024^{***}	0.024^{***}	1.000	1.000
	(0.003)	(0.003)	(0.003)		
N (# exams)	122	122	122		
Panel B. Math/ELA differences in statistics for exam takers on proficiency margin ($\theta_i = \underline{\theta}$)					
Density of ability distribution	0.018^{**}	0.020^{**}	0.017^{**}	0.900	1.000
	(0.009)	(0.009)	(0.009)		
Derivative of expected proficiency	0.186^{***}	0.212^{***}	0.183^{***}	1.000	1.000
	(0.036)	(0.043)	(0.043)		
Density of ability distribution × derivative of expected proficiency	0.087^{***}	0.098^{***}	0.087^{***}	1.000	1.000
	(0.019)	(0.021)	(0.019)		
N (# exams)	122	122	122		
Panel C. Regression for math/ELA ratio of literature estimates (weighted by # exams)					
Math/ELA ratio of test prep incentives	0.834^{***}	0.981^{***}	0.679^{***}	1.000	1.000
	(0.228)	(0.246)	(0.205)		
N (# exams)	63	63	63		
Panel D. Regression for math/ELA ratio of literature estimates (weighted by # lit. estimates)					
Math/ELA ratio of test prep incentives	0.509^{*}	0.690^{**}	0.368	0.300	0.500
	(0.282)	(0.265)	(0.267)		
N (# literature estimates)	19	19	19		

Notes: This table examines the sensitivity of our main results to simulation error. In Panels A–B, column (A) replicates our main estimates for the math/ELA difference in exam statistics (from column B of Table 4). In Panels C–D, column (A) replicates our main estimates from regressing the math/ELA ratio of literature estimates on the math/ELA ratio of test prep incentives (from column A of Table 7). All variables and regression specifications are identical to those described in Tables 4 and 7.

To test the sensitivity of these results, we re-run our simulations 10 times following the same methodology described in Section 4.1 and Appendix C.5. Columns (B)–(E) display results from these 10 replications. Column (B) shows the largest coefficient (in magnitude) from these replications, and column (C) shows the smallest coefficient (in magnitude). Column (D) shows the proportion of coefficients that have a p-value less than 0.05, and column (E) shows the proportion with a p-value less than 0.10.

Parentheses in Panels A–B contain robust standard errors. Parentheses in Panels C–D contain standard errors clustered at the literature estimate level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE A7. Math/ELA estimates in the class size literature

(A)	(B)	(C)	(D)	(E)	(F)	(G)
				Main effects of larger classes		Ratio
Paper	State/country	Years	Grades	Math	ELA	(E)/(F)
Panel A. U.S. papers from pre-accountability years						
McGiverin et al. (1989)	Indiana	1985–86	2	−0.477	−0.461	1.03
Krueger (1999)	Tennessee	1986–89	K–3	−1.960	−3.280	0.60
Hoxby (2000)	Connecticut	1986–98	4,6	−0.098	−0.118	0.84
Akerhielm (1995)	United States	1988	8	−0.020	−0.020	1.00
Average						**0.87**
Panel B. U.S. papers in years with accountability						
Jepsen and Rivkin (2009)	California	1991–02	2–4	−0.100	−0.060	1.67
Rivkin et al. (2005)	Texas	1993–98	4–7	−0.005	−0.004	1.27
Molnar et al. (1999)	Wisconsin	1997–98	1	−0.129	−0.119	1.09
Average						**1.34**
Panel C. International papers (low accountability)						
Angrist and Lavy (1999)	Israel	1991	4–5	−0.140	−0.204	0.69
Urquiola (2006)	Bolivia	1996	3	−0.210	−0.180	1.17
Average						**0.93**
Average (all papers)						**1.04**

Notes: This table shows math and ELA estimates from research on the effects of class size on student achievement. Panel A includes papers with U.S. samples in years prior to the existence of state accountability policies. Panel B includes papers with U.S. samples in years after the introduction of state accountabiltiy policies. In categorizing U.S. papers into Panel A or B, we follow Dee and Jacob (2011)'s classification of the year in which each state implemented a consequential accountability policy (see their Table 1). Panel C includes papers with international samples; in these settings, standardized tests were administered without significant accountability policies.

We include papers that we could find with samples in the 1990s or earlier. We include only papers with identification strategies that are typically perceived as credible by researchers in the economics of education (e.g., random assignment, regression discontinuity, difference in differences, or other instrumental variables). In addition, we include only papers that present estimates on *both* math and ELA scores.

Column (A) lists the papers that meet these criteria. Columns (B)–(D) show the state/country, exam years (spring of academic year), and grades for the sample in each paper. Columns (E)–(F) show math and ELA estimates from the authors' benchmark specification, and column (G) shows the ratio of columns (E) and (F). The units for these estimates vary across papers (e.g., effect sizes, percentiles, and test score growth); in all cases, the math and ELA effects are from the same specification and in the same units. If the authors present estimates for different grades or exams, we report the average of these estimates. We report all estimates as negative numbers because each paper finds that larger class sizes lead to lower test scores. Details on the sources and calculations are available from the authors upon request.

B. Theoretical appendix

B.1. **Optimal test prep.** This appendix presents a full derivation of our framework and proposition from Section 3.

We consider a teacher with a class of students indexed by i. We assume each student's ability is given by

$$\theta_i = \alpha_i + g(e^*)h(\alpha^* - \alpha_i). \tag{B1}$$

The terms are defined as follows:

- α_i = student i's ability prior to any test preparation;
  - For simplicity we assume α_i is continuously distributed with density $f(\alpha_i)$.
- e^* = teacher effort;
- $g(e^*)$ = effect of effort on skill accumulation;
  - We assume $g(0) = 0$, $g'(\cdot) > 0$, and $g''(\cdot) < 0$.
- α^* = teacher's target instruction level;
- $h(\alpha^* - \alpha_i)$ = effect of the distance between α^* and α_i on skill accumulation;
  - We assume $h(\cdot)$ is positive and decreasing as a function of $|\alpha^* - \alpha_i|$, and that $h(\cdot) = 0$ if $|\alpha^* - \alpha_i| > h$ for some $h > 0$.
- θ_i student i's ability after test preparation.

The exam consists of multiple questions indexed by $q = 1, \ldots, Q$. A student's ability affects their exam performance through:

$$p_q(\theta_i) \equiv Pr[u_{iq} = 1|\theta_i], \tag{B2}$$

where u_{iq} is an indicator equal to one if student i correctly answers question q.

We consider test scores defined by whether or not a student meets a proficiency standard:

$$\tau(R_i) = \mathbb{1}\{R_i \geq \underline{\mathrm{R}}\}, \quad \text{where } R_i = \sum_{q=1}^{Q} u_{iq}, \tag{B3}$$

where R_i is student i's *raw score* (i.e., total number of correct answers), and $\underline{\mathrm{R}}$ is the minimum raw score required for proficiency.

The teacher's utility is given by:

$$V(\bar{\tau}, e^*) = \psi\bar{\tau} - c(e^*), \quad \text{where } \bar{\tau} = E[\tau(R_i)]. \tag{B4}$$

$\bar{\tau}$ is the class *proficiency rate*, i.e., the proportion of students in the class who meet the standard of proficiency. $c(e^*)$ is the teacher's cost of effort; we assume $c(0) = 0$, $c'(\cdot) > 0$, and $c''(\cdot) > 0$. The term ψ reflects the relative utility of proficiency and effort. This term could represent the "stakes" of the exam for the teacher, or the utility benefits of test score

gains per unit of effort. This term helps connect our framework to the literatures in Table 1 (see Appendix B.4).

The teacher chooses e^* and α^* to maximize expected utility. Each student has an expected likelihood of achieving proficiency that depends on their post-prep skill, $E[\tau(R_i)|\theta_i]$. The teacher's expected utility is given by the mean expected proficiency across all students minus the cost of effort. Formally, the teacher's problem is:

$$\max_{e^*,\alpha^*} \psi \int_{\alpha_i} E[\tau(R_i)|\theta_i] f(\alpha_i) d\alpha_i - c(e^*), \tag{B5}$$

where the ability of each student, θ_i, depends on α^* and e^*.

The optimal value of α^* is characterized by:[33]

$$\int_{\alpha_i} \frac{dE[\tau(R_i)|\theta_i]}{d\theta_i} h'(\alpha^* - \alpha_i) f(\alpha_i) d\alpha_i = 0. \tag{B6}$$

We have $h'(\cdot) \geq 0$ if $\alpha_i \geq \alpha^*$, and $h'(\cdot) \leq 0$ if $\alpha_i \leq \alpha^*$. Thus (B6) can be written as:

$$\int_{\alpha_i \geq \alpha^*} \frac{dE[\tau(R_i)|\theta_i]}{d\theta_i} h'(\alpha^* - \alpha_i) f(\alpha_i) d\alpha_i = - \int_{\alpha_i < \alpha^*} \frac{dE[\tau(R_i)|\theta_i]}{d\theta_i} h'(\alpha^* - \alpha_i) f(\alpha_i) d\alpha_i.$$

This equation shows that the optimal value of α^* balances the skill gains, $h'(\cdot)$, for students with ability above the target level ($\alpha_i \geq \alpha^*$) with the skill losses, $-h'(\cdot)$, for students below the target level ($\alpha_i < \alpha^*$). These skill gains/losses are weighted by the number of students, $f(\alpha_i)$, and by the effect of an increase in the student's skill on the likelihood of proficiency, $dE[\tau(R_i)|\theta_i]/d\theta_i$. In other words, the optimal value of α^* depends on the distribution of α_i's in the classroom and on the likelihood that an increase in skill, θ_i, will make the difference in terms of achieving proficiency. The optimal α^* will typically be close to the value of α_i for which $E[\tau(R_i)|\alpha_i] = 0.5$, unless the class contains few students at this ability level.

The optimal value of e^* is characterized by:

$$\frac{c'(e^*)}{g'(e^*)} = \psi \int_{\alpha_i} \frac{dE[\tau(R_i)|\theta_i]}{d\theta_i} h(\alpha^* - \alpha_i) f(\alpha_i) d\alpha_i. \tag{B7}$$

All terms in this expression are positive. The lefthand side is increasing in e^* since $c(\cdot)$ is convex and $g(\cdot)$ is concave. The righthand side can be seen as a weighted average of the effect of an increase in the student's skill on the likelihood of proficiency, $dE[\tau(R_i)|\theta_i]/d\theta_i$. The weights are the total amount of skill accumulation for each ability level, α_i, given the teacher's target instruction level, α^*. Total skill accumulation depends on the number of students, $f(\alpha_i)$, and the skill accumulation for each student, $h(\alpha^* - \alpha_i)$. Proposition 1 follows from these observations:

[33] We derive (B6) by taking the derivative of (B5) with respect to α^* using the chain rule. We also remove irrelevant terms; the solution does not depend on the cost of effort, $g(e^*)$, or the exam stakes, ψ.

Proposition 1. *The teacher's optimal level of effort, e^*, is increasing in:*

- *The number of students on the margin of expected proficiency; and*
- *The derivative of expected proficiency with respect to ability, $dE[\tau(R_i)|\theta_i]/d\theta_i$, for marginally-proficient students.*

B.2. **Exam design and proficiency returns to ability.** Proposition 1 shows that teacher effort depends on the derivative of expected proficiency with respect to ability for marginally-proficient students. This term can be written as (see Appendix B.3 for the derivation):

$$\frac{dE[\tau(R_i)|\theta_i]}{d\theta_i} = \sum_{q=1}^{Q} p'_q(\theta_i) E\Big[\tau\Big(R_i^{-q}+1\Big) - \tau\Big(R_i^{-q}\Big)\Big|\theta_i\Big]. \tag{B8}$$

Equation (B8) has two terms. The first term, $p'_q(\theta_i)$, is the effect of an increase in θ_i on the probability of a correct answer to question q. The second term, $E\Big[\tau\Big(R_i^{-q}+1\Big) - \tau\Big(R_i^{-q}\Big)\Big|\theta_i\Big]$ is the *expected proficiency return to question* q, where R_i^{-q} denotes the raw score *excluding* question q. This term represents the effect of a correct answer to question q on the probability that the student achieves proficiency. The total effect of an increase in θ_i on the expected proficiency rate is given by the *sum* (across questions) of the product of these two terms.

In the three-factor IRT model defined by equation (4), the $p'_q(\theta_i)$ term is equal to

$$p'_q(\theta_i) = \frac{a_q(1-c_q)e^{a_q(\theta_i - b_q)}}{(1+e^{a_q(\theta_i-b_q)})^2}. \tag{B9}$$

Equation (B9) shows that, for test takers on the proficiency margin, $dE[\tau(R_i)|\theta_i]/d\theta_i$ is:

- Increasing in the discrimination parameter, a_q;
- Decreasing in the distance between question difficulty and the exam taker's ability, $|\theta_i - b_q|$;
- Decreasing in the psuedo-guessing parameter, c_q.

$dE[\tau(R_i)|\theta_i]/d\theta_i$ is also increasing in the total number of questions, Q, for marginally-proficient students. To see this, suppose all questions are identical, so that equation (B8) can be written as:

$$\frac{dE[\tau(R_i)|\theta_i]}{d\theta_i} = Q p'_q(\theta_i) E\Big[\tau\Big(R_i^{-q}+1\Big) - \tau\Big(R_i^{-q}\Big)\Big|\theta_i\Big]. \tag{B10}$$

Equation (B10) is increasing in Q for exam takers right at the proficiency margin, for whom the proficiency return to a correct answer, $E\Big[\tau\Big(R_i^{-q}+1\Big) - \tau\Big(R_i^{-q}\Big)\Big|\theta_i\Big]$, is large. The intuition is that an exam is more informative about ability when it has more questions, which will more precisely measure increases in ability for marginally-proficient students.

B.3. **Derivation of expression** (B8)**.** The likelihood of proficiency for a student with ability θ_i is:

$$\begin{aligned}
E[\tau(R_i)|\theta_i] &= \sum_{r=0}^{Q} Pr[R_i = r|\theta_i]\tau(r) \\
&= \sum_{r=0}^{Q} \Big\{ Pr[R_i \geq r|\theta_i] - Pr[R_i \geq r+1|\theta_i] \Big\}\tau(r) \\
&= \tau(0) + \sum_{r=1}^{Q} Pr[R_i \geq r|\theta_i] \Big\{ \tau(r) - \tau(r-1) \Big\} \\
&= \tau(0) + \sum_{r=1}^{Q} \Big\{ \sum_{s=r}^{Q} Pr[R_i = s|\theta_i] \Big\} \Big\{ \tau(r) - \tau(r-1) \Big\}
\end{aligned}$$

The derivative of expected proficiency with respect to θ_i is:

$$\begin{aligned}
\frac{dE[\tau(R_i)|\theta_i]}{d\theta_i} &= \sum_{r=1}^{Q} \Big\{ \sum_{s=r}^{Q} \sum_{q=1}^{Q} p'_q(\theta_i) Pr[R_i^{-q} = s-1|\theta_i] - p'_q(\theta_i) Pr[R_i^{-q} = s|\theta_i] \Big\} \Big\{ \tau(r) - \tau(r-1) \Big\} \\
&= \sum_{r=1}^{Q} \Big\{ \sum_{q=1}^{Q} p'_q(\theta_i) Pr[R_i^{-q} = r-1|\theta_i] \Big\} \Big\{ \tau(r) - \tau(r-1) \Big\} \\
&= \sum_{q=1}^{Q} p'_q(\theta_i) \Big\{ \sum_{r=1}^{Q} Pr[R_i^{-q} = r-1|\theta_i] \big\{ \tau(r) - \tau(r-1) \big\} \Big\} \\
&\equiv \sum_{q=1}^{Q} p'_q(\theta_i) E\Big[\tau\Big(R_i^{-q} + 1\Big) - \tau\Big(R_i^{-q}\Big) \Big| \theta_i \Big]
\end{aligned}$$

B.4. **Implications of exam design for estimates in literature.** Equation (B7) shows that teacher effort is increasing in ψ, which reflects the relative utility of proficiency and effort. This term helps relate our framework to the three literatures in Table 1.

Panel A of Table 1 includes papers on accountability systems, which measure the effects of an increase in the exam stakes. Suppose ψ reflects the stakes of the exam for the teacher. Equation (B7) shows than an increase in ψ will lead to a larger change in teacher effort when there are more students on the margin of proficiency, and when the derivative of expected proficiency is larger for marginal students (as in Proposition 1).

Panel B of Table 1 includes papers that estimate the test score effects of attending a charter school. Charter schools may be better than traditional public schools if they have lower costs of effort, or if they are better at converting effort into test score gains. These are equivalent to a higher value of ψ in equation (B7). By the same argument as above, the effects of admission to a school with a higher ψ are increasing in the two factors highlighted by Proposition 1.

The same argument holds for variation in teacher value added, which is the focus of the papers in Panel C of Table 1. The variation between good and bad teachers can be thought of as variation in ψ. If teachers have incentives to engage in test prep, variation in ψ would lead teachers to choose different levels of effort. This variation would be larger when exams have more marginally-proficient students, and when the derivative of the proficiency rate is larger for marginal students.

C. Empirical appendix

This appendix provides details on our data and empirical simulations. We also provide data and codes that execute our simulations and reproduce our results.

C.1. **Data sources and sample.** Our data come from the technical reports of grade 3–8 math and ELA exams in the United States. Appendix Table C1 provides links to each technical report. Panels A–B include reports for the states in our main sample. Panel C includes other states from which we were able to find data for Figure 4.

Our main sample focuses on six states that are the setting for most of the research in Table 1: Florida, Illinois, Massachusetts, New York, North Carolina, and Texas. We collected data on exams in two time periods: 2000–2003 (pre-NCLB) and 2006–2008 (NCLB era). In each time period, we used the earliest exam for which we could find a technical report with all the information necessary for our empirical simulations.

Our simulations rely on three types of information from these reports: 1) question parameters; 2) score scaling; and 3) score distributions. We discuss each of these in turn.

C.2. **Question parameter data.** Our first type of data is information on the question parameters. Figure C1 provides an example of question parameter data for the Massachusetts 2006 grade 8 math exam. These data include the IRT parameters $\{a_q, b_q, c_q\}$ for each question on the exam. Many exams (such as MA-2006) also have open response questions that are worth multiple points. These are parameterized using a partial credit model, which can be converted into an equivalent number of binary questions on the IRT scale.

Columns (A)–(C) in Table C2 show the question parameter data that were available for each exam in our sample. Column (A) shows exams with IRT parameters for each question, as in Figure C1. When these were not available, we used data on the p-values (proportion correct) of each question, as indicated in column (B). We convert these p-values into IRT parameters using the Rasch model.[34] Lastly, column (C) shows exams for which data were only available on the *distribution* of IRT parameters (e.g., the mean and SD of b_q across questions). In this case, we draw questions randomly from a normal distribution with a mean and SD that matches the values from the technical report (see Section C.5 below).

C.3. **Score scaling data.** Our second type of data is information on the scaling of test scores. In most U.S. states, scale scores are a function of raw scores (total correct answers), so all students with the same raw score receive the same scale score. In this case, technical reports often contain "raw-to-scale" conversions, which show the scale score assigned to each potential raw score on the exam. Figure C2 provides an example of the raw-to-scale data

[34] The Rasch model assumes $a_q = 1$ and $c_q = 0$ for all q. We can then solve for question difficulty, b_q, for the average test taker ($\theta_i = 0$).

we use for the Massachusetts 2006 grade 8 math exam. Column (D) in Table C2 shows the exams with raw-to-scale conversions available in the technical reports.

For the exams indicated in column (E) of Table C2, we compute scores using a "theta-to-scale" conversion. The technical reports of these exams include slope and intercept parameters that allow for a linear transformation from expected ability (theta) units to scale score units. In the NC-2001 and NC-2006 exams, scale scores are a function of raw scores, so we compute the expected ability associated with each raw score, $E[\theta_i|R_i]$, in our simulations, and then convert to scale scores using the slope and intercept parameters. In the IL-2000, FL-2003, and FL-2006 exams, scale scores are a function of the full vector of exam responses, $\{u_{i1}, \ldots, u_{iQ}\}$.[35] For these exams, we compute $E[\theta_i|u_{i1}, \ldots, u_{iQ}]$ in our simulations, and use this for the linear transformation to scale scores.

C.4. **Score distribution data.** Our third type of data is information on the distribution of test scores. Figure C3 provides an example of the score distribution data for the Massachusetts 2006 grade 8 math exam. These data show the percentage of test takers who earned each scale score during the test's operation in that year. Column (F) in Table C2 shows exams with distribution data at the scale score level.

For the exams indicated in column (G) of Table C2, information was only available on the percentage of test takers in different performance levels (e.g., Advanced, Proficient, Needs Improvement, Failing). These performance levels are associated with different scale score ranges. Importantly, every technical report includes the minimum scale score necessary to achieve proficiency—a key metric in our analysis.

C.5. **Empirical simulations.** Our empirical simulation for each exam proceeds as follows:

(1) Create $i = 1, \ldots, 1000$ test takers at each ability level $\theta \in \{-5, -4.9, \ldots, 0, \ldots, 4.9, 5\}$.
(2) Use the question parameter data (Section C.2) to draw a random vector of exam responses, $\{u_{i1}, \ldots, u_{iQ}\}$, based on each test taker's ability, θ_i, and the IRT model (equation 4). Each u_{iq} indicates whether test taker i answered question q correctly.
(3) Compute each test taker's raw score, $R_i = \sum_{i=1}^{Q} u_{iq}$.
(4) Compute each test taker's scale score using the score scaling data (Section C.3).
(5) Re-weight the distribution of ability, θ_i, to match the observed score distribution (Section C.4).[36]

[35] In these exams, students with the same raw scores can receive different scale scores if they answered different questions correctly.

[36] Specifically, we begin with the prior that $\theta_i \sim \phi(\theta_i)$, where $\phi(\cdot)$ is the standard normal density. We compute the fraction of test takers at each scale score s in the simulated data, and denote this by P_s^{ϕ}. Lastly, we compute the density of the realized ability distribution, $f_s(\theta_i) = \phi(\theta_i)P_s/P_s^{\phi}$ for each score s, where P_s is the realized score distribution (Section C.4). We collapse the data to the θ level to compute the density of the ability distribution at each value of $\theta \in \{-5, -4.9, \ldots, 0, \ldots, 4.9, 5\}$, i.e., $f(\theta) = E[f_s(\theta_i)|\theta_i = \theta]$.

(6) Compute the statistics for the main results in Tables 2–3, which include:

- Proficiency rate and proportion correct over all exam takers;
- Level of ability, $\underline{\theta}$, at which Pr(Proficient) = 0.5 ("proficiency margin");
- Density of ability distribution at proficiency margin;
- Derivative of expected proficiency at the proficiency margin (see equation (B8)).

We provide simulation codes for each exam in our main sample.

Item	Type	A	B	C	D1	D2	D3	D4
226896	MC	1.379	0.343	0.138				
226908	MC	1.406	0.052	0.158				
226910	MC	0.976	-0.573	0.091				
226940	MC	1.487	-0.194	0.277				
226944	MC	1.017	-0.665	0.103				
226983	MC	1.507	-0.738	0.261				
226996	MC	0.761	-0.193	0.294				
228708	MC	0.739	0.279	0.421				
228779	MC	0.680	-0.312	0.201				
228816	MC	0.962	-0.241	0.140				
228847	MC	1.686	-0.558	0.259				
229496	MC	0.995	-0.064	0.199				
229502	MC	1.269	0.219	0.225				
229588	MC	1.043	-0.370	0.443				
229608	MC	1.403	-0.776	0.098				
229613	MC	1.091	-0.313	0.101				
229673	MC	1.439	0.636	0.172				
229680	MC	0.875	-0.585	0.164				
229702	MC	0.983	0.829	0.347				
229722	MC	1.486	0.839	0.071				
229724	MC	0.668	-1.735	0.095				
248137	MC	0.596	1.101	0.192				
248139	MC	1.097	0.018	0.302				
248140	MC	1.444	-0.294	0.309				
248142	MC	0.889	0.330	0.166				
248144	MC	1.110	0.462	0.292				
248155	MC	1.120	-1.065	0.098				
248162	MC	0.552	0.108	0.092				
248168	MC	0.577	-1.565	0.000				
228756	SA	1.139	0.374					
228958	SA	0.944	0.356					
229539	SA	0.955	-0.250					
229619	SA	0.962	-1.254					
229706	SA	0.973	0.147					
227751	OR	1.176	-0.321		1.473	0.517	-0.576	-1.414
228825	OR	0.869	-0.060		0.957	0.490	-0.302	-1.145
248147	OR	1.042	-0.712		1.077	0.227	-0.210	-1.094
248152	OR	1.333	0.086		1.304	0.393	-0.625	-1.072
248173	OR	1.064	-0.827		0.967	0.416	-0.181	-1.202

FIGURE C1. Massachusetts grade 8 math (2006) — Question parameter data

Source: 2006 Massachusetts (MCAS) Technical Report, Appendix B, p. 13. Available in October 2021 at: http://www.mcasservicecenter.com/documents/MA/Technical%20Report/TechReport_2006.htm.

Grade 8 Raw Score	Mathematics Scaled Score
0	200
1	200
2	202
3	202
4	202
5	204
6	204
7	206
8	208
9	210
10	210
11	212
12	212
13	212
14	214
15	214
16	214
17	216
18	216
19	216
20	216
21	218
22	218
23	218
24	218
25	218
26	220
27	220
28	220
29	220
30	222
31	224
32	226
33	228
34	230
35	232
36	234
37	236
38	238
39	240
40	240
41	242

Grade 8 Raw Score	Mathematics Scaled Score
42	244
43	246
44	248
45	250
46	252
47	256
48	258
49	260
50	262
51	264
52	268
53	274
54	280

FIGURE C2. Massachusetts grade 8 math (2006) — Score scaling data

Source: 2006 Massachusetts (MCAS) Technical Report, Appendix C, pp. 9–10. Available in October 2021 at: http://www.mcasservicecenter.com/documents/MA/Technical%20Report/TechReport_2006.htm.

Score	Number	Percentage	Cumulative Percentage
200	106	0.14	0.14
202	160	0.21	0.35
204	481	0.64	1.00
206	455	0.61	1.60
208	572	0.76	2.36
210	1406	1.87	4.24
212	2643	3.52	7.76
214	3128	4.17	11.93
216	4617	6.15	18.09
218	7261	9.68	27.76
220	6645	8.86	36.62
222	1780	2.37	38.99
224	1832	2.44	41.44
226	1897	2.53	43.97
228	1834	2.44	46.41
230	1858	2.48	48.89
232	1980	2.64	51.53
234	1944	2.59	54.12
236	1997	2.66	56.78
238	1976	2.63	59.41
240	3991	5.32	64.73
242	2053	2.74	67.47
244	2048	2.73	70.20
246	2096	2.79	72.99
248	2219	2.96	75.95
250	2166	2.89	78.84
252	2175	2.90	81.74
256	2212	2.95	84.69
258	2171	2.89	87.58
260	2258	3.01	90.59
262	2059	2.74	93.33
264	1855	2.47	95.81
268	1555	2.07	97.88
274	1064	1.42	99.30
280	527	0.70	100.00

FIGURE C3. Massachusetts grade 8 math (2006) — Score distribution data

Source: 2006 Massachusetts (MCAS) Technical Report, p. 120. Available in October 2021 at: http://www.mcasservicecenter.com/documents/MA/Technical%20Report/TechReport_2006.htm.

TABLE C1. Sources for technical reports

State	Year	Archive .org?	URL
Panel A. Pre-NCLB exams (2000–2003)			
FL	2003		http://www.fldoe.org/accountability/assessments/k-12-student-assessment/archive/fcat/researchers.stml
IL	2000		https://www.isbe.net/Pages/Illinois-Standards-Achievement-Test-(ISAT)-Archive.aspx
MA	2000	✓	http://www.doe.mass.edu/mcas/tech/?section=techreports
NC	2001	✓	http://www.ncpublicschools.org/accountability/testing/technicalnotes
NY	2002	✓	http://www.p12.nysed.gov/assessment/reports/archive3.html
TX	2001	✓	http://www.tea.state.tx.us/student.assessment/reporting/
Panel B. NCLB era exams (2006–2008)			
FL	2006		http://www.fldoe.org/accountability/assessments/k-12-student-assessment/archive/fcat/researchers.stml
IL	2008		https://www.isbe.net/Pages/Illinois-Standards-Achievement-Test-(ISAT)-Archive.aspx
MA	2006		http://www.mcasservicecenter.com/documents/MA/Technical%20Report/TechReport_2006.htm
NC	2006	✓	http://www.ncpublicschools.org/accountability/testing/technicalnotes
NY	2006	✓	http://www.p12.nysed.gov/assessment/reports/archive3.html
TX	2006	✓	http://www.tea.state.tx.us/student.assessment/reporting/
Panel C. Other states in Figure 4 (2006–2009)			
AZ	2009		https://cms.azed.gov/home/GetDocumentFile?id=5852b7edaadebe0658611ca8
CA	2006		https://star.cde.ca.gov/startechnicalreports.asp
CO	2007		https://www.cde.state.co.us/assessment/coassess-additionalresources
LA	2006		https://www.louisianabelieves.com/resources/contact-us (email request)
MD	2006	✓	http://archives.marylandpublicschools.org/msde/divisions/planningresultstest/2006+MSA+Reading+Technical+Report.htm
MO	2006		https://dese.mo.gov/college-career-readiness/assessment/assessment-technical-support-materials
NH	2006		https://www.ride.ri.gov/Portals/0/Uploads/Documents/Instruction-and-Assessment-World-Class-Standards/Assessment/NECAP/TechnicalReports/2006-07-NECAP-Math-Reading-Writing-Technical-Report-with-Appendices.pdf
NJ	2009	✓	https://www.nj.gov/education/assessment/es/njask_tech_report09.pdf
NM	2006		https://files.eric.ed.gov/fulltext/ED500392.pdf
OH	2006		http://education.ohio.gov/getattachment/Topics/Testing/Testing-Analysis-and-Statistics/Statistical-Summaries-and-Item-Analysis-Reports/March-2006-Grades-3-8-OAT-Statistical-Summary.pdf.aspx
PA	2007	✓	https://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/Technical%20Reports/2007%20Reading%20and%20Mathematics%20PSSA%20Technical%20Report.pdf
SD	2008		https://doe.sd.gov/assessment/documents/DS08TRepr.pdf
WA	2006		https://www.k12.wa.us/student-success/assessments/state-testing-overview/scores-and-reports/testing-statistics-frequency-distribution

Notes: This table provides links to the technical reports that contain the data for this paper. Panels A–B include exams for the states/years included in our main analysis. Panel C includes states with data that we use in Figure 4. Entries with a checkmark in the third column were accessed using the Wayback Machine at archive.org. NH represents the New England Common Test Program (NECAP) exam, which was administered in New Hampshire, Rhode Island, and Vermont.

TABLE C2. Data on exam design

			(A)	(B)	(C)	(D)	(E)	(F)	(G)
			Question parameters			Score scaling data		Score distribution data	
State	Year	Grades	IRT parameters	p-values	IRT parameter means/SDs	Raw to scale	Theta to scale	Scale score distribution	Performance level distribution
Panel A. Pre-NCLB exams (2000–2003)									
FL	2003	3–8			✓		✓		✓
IL	2000	3,5,8	✓				✓		✓
MA	2000	4,8	✓			✓			✓
NC	2001	3–8			✓		✓		✓
NY	2002	4,8		✓		✓		✓	
TX	2001	3–8		✓		✓		✓	
Panel B. NCLB era exams (2006–2008)									
FL	2006	3–8	✓				✓		✓
IL	2008	3–8			✓	✓			✓
MA	2006	3–8	✓			✓		✓	
NC	2006	3–8			✓		✓		✓
NY	2006	3–8	✓			✓		✓	
TX	2006	3–8		✓		✓		✓	

Notes: This table shows the data we obtained from the technical reports for each of the six states in our main sample. Panel A includes exams in the pre-NCLB era (2000–2003), and Panel B includes exams in the NCLB era (2006–2008). Columns (A)–(C) show the data included in the technical report related to the question parameters (e.g., difficulty, discrimination, and guessability). Columns (D)–(E) show the data that we use to convert from simulated raw scores to scale scores. Columns (F)–(G) show the type of data available in each report on the realized distribution of test scores. See Appendix Sections C.2–C.4 for details on each type of data.