Incentive Design in Education: An Empirical Approach*

by Hugh Macartney, Robert McMillan, and Uros Petronijevic†

September 2017

Abstract

While incentive schemes to elicit greater effort in organizations are widespread, the effort-incentive strength relationship is difficult to ascertain in practice, hindering incentive design. We propose a new semi-parametric method for uncovering this relationship in an education context, using exogenous incentive variation and rich administrative data. We then devise a structural estimation procedure that allows us to recover the primitives underlying the effort function, based on a model of effort setting. The parameter estimates combined with the model form the basis of a counterfactual approach for tracing the effects of various accountability systems on the *full* distribution of scores for the first time. We show higher average performance comes with greater inequality of outcomes for widespread fixed-target schemes, and that incentive designs not yet enacted can, at no extra cost, improve student performance while reducing test score inequality – of relevance to education reform.

Keywords: Incentive Design, Effort, Accountability Scheme, Education Production, Semi-Parametric, Counterfactual, Test Score Distribution, Inequality, Education Reform

JEL Classifications: D82, I21, J33, M52

*We would like to thank Joseph Altonji, Peter Arcidiacono, David Deming, Giacomo De Giorgi, David Figlio, Caroline Hoxby, Lisa Kahn, Lance Lochner, Rich Romano, Eduardo Souza-Rodrigues, Aloysius Siow, and seminar participants at Duke University, the University of Florida, the NBER, SITE, Western, and Yale University for helpful comments and suggestions. Thanks also to Hammad Shaikh and Kaili Yang for excellent research assistance. Financial support from SSHRC and the University of Toronto Mississauga is gratefully acknowledged. All remaining errors are our own.

†Contact information: Macartney – Duke University and NBER, hugh.macartney@duke.edu; McMillan – University of Toronto and NBER, mcmillan@chass.utoronto.ca; Petronijevic – York University, upetroni@yorku.ca.

I. Introduction

Across many types of organization, schemes that provide incentives to exert effort are often seen as an important means of boosting organizational performance. The design of such schemes has, naturally, been a central preoccupation in economics and also a very challenging one, given that effort is typically unobserved. While this challenge has been taken up in a substantial body of sophisticated research in contract theory,[2] a host of incentive schemes operating in practice are only loosely informed by the theoretical contracting literature. This creates scope – in a variety of settings – for potentially significant performance gains from judicious incentive reform.

In order to gauge whether such gains are attainable, one approach involves studying the introduction of actual incentive reforms, as in classic papers by Lazear (2000) and Bandiera, Barankay and Rasul (2005), for example.[3] Incentive designers often wonder about more speculative considerations, however, looking to the effects of changing the parameters of existing schemes counterfactually, or the effects of incentive schemes yet to be implemented in practice. As a complement to the evaluation of actual schemes, therefore, approaches that combine a strategy for identifying effort under prevailing incentive provisions with a framework for counterfactual analysis are especially appealing.[4]

In this paper, we propose such an approach, consisting of three related parts. First, we outline a new semi-parametric method for recovering the relationship be-

[2]See the seminal work of Mirrlees (1975) and subsequent theoretical analyses: the robustness of optimal contract forms has been an important theme in the contracting literature.

[3]Lazear's well-known study shows how the introduction of a new piece-rate style incentive scheme by Safelite Glass Corporation led to an increase in company profits, implying that the pre-existing contract was suboptimal. Bandiera *et al.* demonstrate that significant productivity gains arise among fruit pickers, moving to a piece rate from a relative incentive scheme. Other papers consider incentive variation more broadly, including Mas and Moretti's innovative 2009 study of the productivity effects of varying peers among supermarket checkout staff.

[4]A recent body of innovative research studying worker incentives adopts this kind of approach – see important papers by Copeland and Monnet (2009) and Misra and Nair (2011), among others.

tween effort and incentive strength. Second, we estimate the primitives underlying this effort function based on a structural model of effort setting. Third, the parameter estimates combined with the model then form the basis of a simulation approach, which we use to explore how changing the features of incentive schemes counterfactually affects relevant outcomes.[5]

We develop this combined approach in the context of accountability systems in public education – a prominent policy arena in which incentive schemes have been adopted widely. In their essence, such schemes involve setting performance targets and explicit rewards (or penalties) that depend on target attainment, their goal being to increase teacher and school effort, in turn raising test scores.[6] Several types of accountability scheme have been implemented to date, including proficiency schemes – notably the federal No Child Left Behind Act of 2001 ('NCLB') – that set fixed performance targets based on school sociodemographics, and value-added schemes whose targets condition on prior student scores.

Such variety brings to mind important incentive design issues, particularly how the features of different accountability systems affect student and school outcomes throughout the performance distribution. These types of consideration are particularly salient, given that states have been revisiting incentives under NCLB. While several convincing studies consider incentive issues by focusing on particular aspects of accountability schemes already in operation,[7] our approach allows us to analyze the impacts of alternative education accountability systems, including ones not yet

[5]We take a positive approach, contrasting with the normative emphasis of the optimal contracting literature. Rather than writing down the planner's objective and deriving the optimal contract form given the relevant constraints, we start out by recovering the effort distribution under prevailing incentives (described below). Then we use the estimated primitives of the effort-setting problem to explore how the distribution of accountability scheme outcomes might change, and be improved, by varying incentives.

[6]Persuasive evidence that accountability schemes succeed in improving student achievement already exists – see Carnoy and Loeb (2002), Burgess *et al.* (2005), Lavy (2009), Hanushek and Raymond (2005), Dee and Jacob (2011), and Imberman and Lovenheim (2015), among others. Figlio and Kenny (2007) offer an interesting alternative perspective: in their study, achievement gains could be due to better schools being the ones that adopt stronger incentives for staff.

[7]See Cullen and Reback (2006), Neal and Schanzenbach (2010), and Macartney (2016).

implemented, across the entire distribution of test scores. In so doing, we are able to assess – for the first time – the relative merits of rival schemes in a quantifiable manner, and shed light on the way accountability incentives affect educational outcome inequality, relevant to the broader inequality debate.

Our semi-parametric approach for recovering the relationship between effort and incentive strength – the paper's first main contribution – exploits plausibly exogenous incentive variation arising from the introduction of NCLB. Being a 'fixed' scheme, NCLB creates incentives to focus on students at the margin of passing relative to a fixed target – a feature that has been well-documented in the literature (see Reback (2008), for example). We take advantage of this non-uniformity in North Carolina, a setting for which we have rich administrative data covering all public school students over a number of years.

To guide the approach, one can think of a model of the education process that links incentives to outcomes via discretionary actions – 'effort' – which we will take to refer to changes in observable test scores attributable to incentive variation. In general, such a model implies an effort function (solving for optimal effort) that depends on the parameters of the incentive scheme and, under threshold targets, a measure of incentive strength that captures how close to the target a student is predicted to be.

We estimate this function semi-parametrically. First, we construct a continuous incentive strength measure for each student using rich data from the pre-NCLB-reform period, equal to the gap between the target and the student's predicted score; this describes how marginal each student is. Then we compare the achievement of each student against a prediction reflecting all pre-reform inputs; this difference for each level of incentive strength serves as a pre-period performance control. Once incentives are altered following NCLB's introduction, teachers and schools re-optimize, and the post-reform difference between the realized and predicted test scores will reflect both the original inputs as well as any additional

effort associated with the new non-uniform incentives, likely to be strongest where students are marginal.[8]

Consistent with standard intuition, we find that the profile of actual scores in the pre-reform period plotted against the incentive measure is remarkably flat. Then, once the reform comes in, there is a pronounced hump, peaking precisely where incentives should be most intense and declining on either side of that.[9] By differencing the post- and pre-reform distributions, we can then uncover – based on minimal assumptions[10] – the underlying effort response to greater accountability for all levels of incentive strength.

The paper's second main contribution is to set out and estimate a structural model of effort setting that allows us to estimate the primitives underlying this effort function. The advantage of the model is that it makes explicit how effort responds when the parameters of the incentive scheme change. It turns out we are able to identify the parameters of the structural model in a credible way, and the estimates are new and plausible.

The paper's third contribution is to develop a simulation framework based on the model and estimates, which we use to conduct informative counterfactuals that provide the first quantitative analysis of the relative merits of alternative incentive systems in education, including schemes yet to be enacted. The approach allows us to recover the full counterfactual outcome distribution under various accountability schemes, having placed them all on a common footing by equating costs.

[8]This is related to the approach in Neal and Schanzenbach (2010). In their analysis, they only have one year of pre-reform data, so do not construct a score difference relative to the pre-reform. They also focus on deciles of the distribution, while we develop a *continuous* measure of incentive strength, important for our counterfactuals.

[9]We also find evidence supporting the hypothesized channel, rather than rival stories involving schools adjusting other relevant inputs, or focusing on the middle of the distribution.

[10]We assume first that the education production technology is linear in effort and separable – a reasonable first-order approximation, made almost without exception in the education literature. (Below, we discuss how we are able to assess the extent of non-linearities.) Second, NCLB should influence the effort decisions of educators but not the other determinants of student test scores – an assumption we are able to check indirectly.

Three main findings emerge, each of which is relevant to incentive design in education. First, fixed targets (of the form taken by NCLB) give rise to a clear, quantitatively significant tradeoff between the average effort exerted by teachers and test score inequality across students.[11] Second, student-specific *bonuses* improve the performance of standard fixed target regimes significantly: attaching higher weight (in the form of bonus payments) to low-performing students raises mean effort by 0.05 standard deviations of the test score and reduces the test score variance by 18 percent. It also reduces the black-white test score gap by 11 percent and the score gap between children of college educated versus non-college educated parents by 10 percent, for no extra resources. Third, student-specific *targets* allow policymakers to reduce unequal treatment of students without sacrificing aggregate effort. We show that switching from fixed to VA targets reduces inequality in teacher effort across students by as much as 80 percent, with at least as much mean effort.

The rest of the paper is organized as follows: In the next section, we provide some basic intuition behind our approach for uncovering the effort response to incentives. In Section III, we describe the exogenous incentive variation and rich administrative data from the state of North Carolina we use in the empirics. Section IV describes our semi-parametric approach for uncovering the effort-incentive strength relationship: we provide results from our implementation of this approach in Section V, along with semi-parametric estimates of the effort function and evidence that motivates our structural model. Next, we set out an estimable model of effort setting in Section VI. The estimation and identification of the structural parameters are discussed in Section VII, followed by the parameter estimates and model fit in Section VIII. Section IX describes our counterfactual framework, and in Section X, we present the main results from the counterfactual analysis. Section XI concludes.

[11] Moving the proficiency target from the 5th to the 95th percentile of the predicted score distribution increases teacher effort by 0.28 standard deviations of the test score, while increasing the test score variance by a factor of three.

II. Recovering the Effort Function – the Basic Idea

Our main goal is to understand how incentive policies affect effort. As effort is typically unobserved, the approach we develop below uses the fact that, as a productive input, effort should be reflected in observed output (test scores). Specifically, we draw on the well established prediction regarding threshold shemes from prior work – see Reback (2008), for instance – that they should create non-linear incentives to apply effort. Thus their introduction should give rise to a corresponding non-linear change in test scores.

To explain the relevant shape prediction and to motivate aspects of our semi-parametric approach, we provide a simple illustrative example.[12]

Threshold Incentives: An Example

Consider a teacher who teaches three students, ordered by their underlying ability. (Thus, the first student has low ability, the second, medium ability, and the third, high ability.) Student ability is an input into education production, the higher is student ability, the higher a student will score on average, allowing test performance to have a random noise component.

Teacher effort is also a productive input that helps raise student scores. Suppose that teacher effort can be directed to an individual student, boosting that student's specific chance of passing. Effort is costly to provide, however. Thus the teacher will balance the likely gains from directing extra effort to individual students against the extra costs.

Our interest is in the way accountability incentives affect effort choices. Consider a threshold scheme – widespread in an education setting – which sets a passing score. Under such a scheme, the benefit of effort is non-linear, being highest around the passing threshold. Intuitively, effort will be most productive for marginal stu-

[12] Appendix A provides a stylized model and an illustrative figure corresponding to the discussion here.

dents – those close to the threshold – allowing them to satisfy the performance target and thus yield an associated accountability benefit for the teacher.

In this example, suppose the passing threshold is set closest to the middle student. The marginal benefit of effort will then be highest for that student, as an extra increment of effort will be more productive than for the other two students, in the sense that it is most likely to move the middle student over the passing threshold. In contrast, the low-ability student may be a 'lost cause,' for whom no amount of effort would lead the student to pass, while the high-ability student will pass even if the teacher ignores that student.

Optimal student-specific effort will be found where the marginal benefit of effort equals the marginal cost. It is straightforward to see that if marginal cost is upward-sloping (for instance, if the effort cost is quadratic), then optimal effort will be highest for the middle-ability student – the one closest to the passing threshold.

Three General Features

Three features of the example are relevant to the semi-parametric approach we develop. First, the example suggests a natural measure of *incentive strength* – one which we use below. This is a student-specific measure that is the difference between the student's predicted score ($\hat{y}$), based on all information prior to the accountability reform being in place and the threshold ($\bar{y}$). Because it will play such an important role in what follows, we will label this *ex ante* measure $\pi \equiv \hat{y} - \bar{y}$. To be clear, it is a *continuous* measure that can be constructed for each student, given appropriate data.

Second, the solution to the teacher's problem yields an effort function that relates the underlying 'production' conditions – in this case, student ability (captured by the student's predicted score) and features of the accountability scheme (the position of the threshold) – to a corresponding effort level. Those conditions can be summarized, in the simplest instance, by our incentive strength measure $\pi \equiv \hat{y} - \bar{y}$,

the difference between the target and the predicted score.

Third, the effort function should follow an inverted-U relationship between incentive strength (just defined) and effort under a threshold scheme. In the example just rehearsed, effort will be low for the low-ability student, high for the middle student, and low for the high-ability student: more generally, the function should peak close to the threshold and decline on either side.[13]

In practice, the precise shape of $e(\pi)$ is unknown. Our semi-parametric approach provides a transparent way of uncovering it.[14]

III. Institutional Setting and Data

Given the first goal of our study – to measure the relationship between effort and accountability incentives – we need exogenous incentive variation and rich administrative data. The state of North Carolina provides both.

On the incentive front, we make use of the introduction of NCLB provisions in North Carolina in the 2002-03 school year, following the passage of the federal No Child Left Behind Act in 2001. NCLB focuses on penalties for under-performing schools, the aim of the program being to close performance gaps by requiring schools to meet Adequate Yearly Progress ('AYP') targets for all students and for each of nine student subgroups. We focus on AYP targets for all students as a reasonable approximation to the prevailing incentives, abstracting from the subgroup aspect in our analysis.[15]

[13] In addition to the stylized features of the example already described, a unimodal error centred around zero will be sufficient for this result. See Appendix A.

[14] A further issue relates to what the resulting shape says about the primitives underlying the effort decision. We show that the those can be backed out under some plausible additional structure. To that end, in Section VI we write down a structural model of teacher effort setting, which we then take to the data.

[15] The state also provided pre-existing pecuniary incentives under the ABCs of Public Education legislation, which applied to all schools serving kindergarten through grade eight starting in the 1996-97 school year. Under the ABCs, each grade from three to eight in every school was assigned a school-grade-specific target gain, depending on both average prior student performance and a constant level of expected test score growth. The ABCs pays a monetary bonus to all teachers

In addition to this incentive variation, North Carolina offers incredibly rich longitudinal education data from the entire state, provided by the North Carolina Education Research Data Center (NCERDC). These data contain yearly standardized test scores for each student in grades three through eight and encrypted identifiers for students and teachers, as well as unencrypted school identifiers. Thus students can be tracked longitudinally, and linked to a teacher and school in any given year.

Our sample period runs from 1997-2005. To focus on schools facing similar incentives, we limit the sample to schools serving kindergarten to eighth grade, and exclude vocational, special education, and alternative schools.[16] These restrictions notwithstanding, our sample sizes are very large, with over five million student-grade-year observations over the nine-year window, and over 14,000 school-year observations.

Table F.1 provides sample summary statistics. Our main performance variables are constructed from individual student test scores. These are measured on a developmental scale, designed so that each additional point represents the same amount of knowledge gained, irrespective of the baseline score or school grade. Both the mathematics and reading scores in the table show a monotonic increase across grades, consistent with knowledge being accumulated in those subjects over time. The test score *levels* are relevant under NCLB, which requires that each student exceeds a target score on standardized tests (among other requirements). The longitudinal nature of the data set also allows us to construct growth score measures for both mathematics and reading, based on within-student gains. These gains are positive, on average, the largest gains occurring in the earlier grades.

With respect to the school-level performance variables, 37 percent failed NCLB

and the principal if a school achieves its overall growth target, based on average school-level gains across all grades.

[16]We also only retain schools with a highest grade served between grades five and eight, thus avoiding the different accountability provisions that arise in high schools as well as the potentially different incentives in schools with only one or two high-stakes grades.

across the sample period.[17] Throughout our sample period, the average school had a school-wide proficiency rate of 79 percent on math and reading tests.

IV. Semi-Parametric Approach

The estimation approach presented in this section is central to our analysis. Our goal is to uncover the optimal effort response $e^*(\pi)$ for a given incentive strength π, introduced in Section II. The strategy we follow makes use (as noted) of the new performance requirements under NCLB as an exogenous shock to the effort decision-making process of educators in 2003.

In order to explain the approach, we lay out the technological assumptions we are making and describe the construction of our *ex ante* incentive strength measure. Then we show how double-differencing combined with an exogeneity argument yields the effort response to incentives. (As will be clear, the approach rests on minimal assumptions.)

Technology: We specify a simple linear structure for the test score production function. This is standard in the literature, and also serves as a reasonable first-order approximation to a richer underlying test score technology.[18]

We think of there being a 'pre-reform' environment in which effort is approximately uniform, irrespective of incentive strength π. Test scores in this environment are generated according to $y(\pi) = \hat{y} + \epsilon(\pi)$, the sum of a systematic component, which may include baseline effort, and noise. We reference a particular score by our *ex ante* incentive measure π, as we are interested in seeing how changes in formal incentives are reflected in the score distribution in a way attributable to an effort response.

To that end, consider a reform R that introduces new performance targets for

[17] Recall that NCLB is a proficiency count system, which assesses school performance according to the fraction of students achieving proficiency status on End-of-Grade tests.

[18] We consider complementarities in the counterfactual analysis below.

teachers, thereby changing the incentives to exert effort. The targets can be written $y_R^T(\hat{y}_R)$, where $\hat{y}_R$ represents the predicted score in the post-reform environment, excluding any additional effort response e^* to the reform. We will write scores in this post-reform environment, using the linearity assumption, as

$$y_R(\pi) = \hat{y}_R + e^*(\pi) + \epsilon_R(\pi), \tag{1}$$

expressed as a function of π.

Incentive Strength and Effort Response: We capture the optimal effort response as a function of incentive strength using a straightforward procedure, in which we distinguish 2003 – the year in which the new incentives came into effect – from pre-reform years. This involves the following steps: First, we predict student performance in a flexible way in those pre-reform years using several covariates, including lagged test scores.[19]

Second, we use the saved coefficients from the first step to construct our *ex ante* incentive strength measure. In particular, combining those coefficients with updated covariates from 2003 and prior test scores for 2002, we are able to predict performance ($\hat{y}$) for 2003. Using the known NCLB target specified by the reform (y^T), we then compute our continuous measure of incentive strength as the difference between the predicted value (which does not include *additional* effort in 2003) and the target: $\pi \equiv \hat{y} - y^T$. On this basis, the predicted score component is invariant to any changes occurring in 2003. Instead, variation in incentive strength when new incentives are considered arises from changes in the target. Specifically, the proficiency target y^T becomes relevant under NCLB, implying that π will capture the strength of effort incentives in 2003 but not in prior years. In the Appendix, we present evidence supporting the view that the shock to incentives was indeed

[19]Specifically, we regress contemporaneous 2002 scores on cubics in prior 2001 math and reading scores and indicators for parental education, gender, race, free or reduced-price lunch eligibility, and limited English proficiency.

exogenous by demonstrating that there is no 'bunching' of the predicted score ($\hat{y}$) distribution around the target (y^T).

Third, having constructed the continuous *ex ante* incentive strength measure, we then turn to the main task – determining the effort response for each value of this continuous incentive strength measure (π). We do this in the following semi-parametric way: for each value of π, we compute the difference between the realized and predicted scores ($y - \hat{y}$) in 2003, the year when the incentive shock occurred. Intuitively, this quantity should contain the effort response as well as any noise in our prediction.[20] In terms of the above technology, this step will recover $y_R(\pi) - \hat{y}_R = e^*(\pi) + \epsilon_R(\pi)$. Recall that the response to NCLB is predicted to be non-uniform (larger for marginal students with π close to zero and smaller for non-marginal students with larger absolute values of π). This prediction serves as a helpful check that our assumptions are satisfied.

A fourth step is designed to control for any pre-existing patterns that are common across the pre- and post-reform periods. To that end, we repeat steps one through three using only the pre-reform years 1998 through 2000.[21] That is, we regress 1999 scores on cubics in prior 1998 math and reading scores as well as contemporaneous student covariates. Using the resulting coefficients, we construct $\hat{y}$ and, combined with the target y^T, π for the year 2000. We then compute $y - \hat{y}$ in 2000 for each value of the incentive measure. This fourth step thus recovers the noise in the pre-reform period ($y(\pi) - \hat{y} = \epsilon(\pi)$).

Our fifth and final step differences the post- and pre-reform distribution from step three and four, to identify the effort function. The double differencing yields:

$$\big(y_R(\pi) - \hat{y}_R\big) - \big(y(\pi) - \hat{y}\big) = e^*(\pi) + \epsilon_R(\pi) - \epsilon(\pi)\,. \tag{2}$$

[20] Notably, the effort response is in addition to the effort exerted under the pre-existing value-added ABCs scheme, even if is not completely uniform. We only require that the ABCs scheme affects y and $\hat{y}$ in the same way for this to be true.

[21] We select these pre-reform years since the test scores in them are all obtained from the same first edition testing suite.

In our context, an exogeneity assumption implies that the RHS of (2) is just equal to $e^*(\pi)$, the desired object. That is, conditional on π, the stochastic components of the production technology are equal in expectation over time. Given that NCLB should influence the effort decisions of educators but not the other determinants of student test scores, this assumption is plausible – recall that the targets under NCLB are student-invariant.[22] We consider supportive evidence next.

V. Incentive Response Results

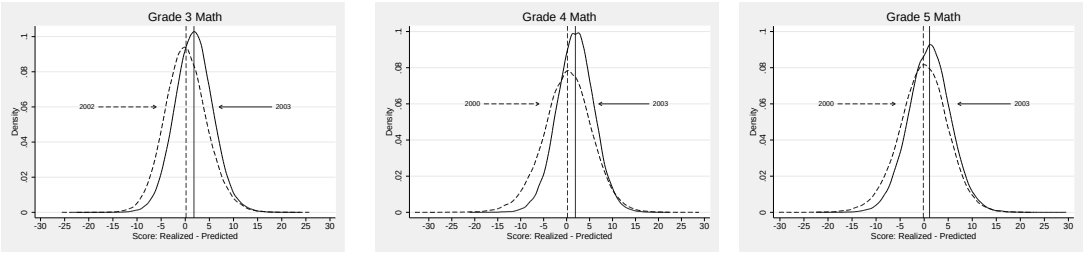
In this section, we present results from the implementation of our semi-parametric research design, and discuss evidence relating to its validity.

V.A. The Test Score Response

Using our rich test score microdata, we can compute whether there was any response to the introduction of NCLB in 2003 based on raw test scores. Figure 1 shows the densities of realized minus predicted test scores in both the pre-period (2000 and 2002)[23] and the post-period (2003), which we interpret as the densities of unobservable test score determinants, including the effort of educators. Predicted scores represent the test scores that are likely to occur in a given year if the relationship between student observable characteristics and realized test scores remains the same as it was in past years. The difference between realized and predicted scores in the pre-NCLB year is centered approximately around zero, suggesting that the prediction algorithm performs well. In 2003, however, the residual densities for all grades display clear rightward shifts, indicating that realized scores exceeded pre-

[22]Note that this strategy allows for the existence of some non-uniformity in the pre-existing value-added ABCs scheme, which is time-invariant.

[23]For grades four and five, we use 2000 as the pre-reform year, rather than the year immediately preceding the implementation of NCLB (2002). North Carolina altered the scale used to measure end-of-grade math results in 2001, implying that contemporaneous and prior scores are on different scales in that year, preventing the use of our prediction algorithm in 2002. We do use 2002 as the pre-reform year in grade three, because these students write the 'pre' test at the beginning of the year, meaning that both scores are still on the same scale in 2001.



(a) 3rd Grade (b) 4th Grade (c) 5th Grade

FIGURE 1 – DENSITIES OF REALIZED MINUS PREDICTED SCORES (2003 VERSUS 2000 OR 2002)

dicted scores on average. This observation is consistent with an improvement in some unobserved determinant of test scores.

V.B. The Estimated Effort Function

Our claim is that the unobserved determinant in question is teacher effort, drawing on the intuitive (and well-established) theoretical predictions associated with proficiency-based accountability schemes, discussed in Section II.

Such schemes reward schools (or refrain from punishing them) based on the percentage of proficient students and so provide clear incentives to focus on students predicted to score around the proficiency target. Students likely to score far below the target require a prohibitively costly amount of extra effort to reach proficiency status, while students predicted to score far above the target are likely to pass without any additional effort at all. Thus, to the extent that the documented shifts in residual densities represent an effort response, we should see the largest gains in realized-over-predicted scores for the students predicted to score near the proficiency threshold.

Figure 2 shows that these are exactly the patterns we find across the predicted test score distribution.[24] In 2003, the gains above predicted scores are low for students predicted to be far below the proficiency threshold; they begin to increase

[24]We focus on the fourth grade distribution in our analysis.

for students predicted to be close to the threshold; and they decline again for students predicted to be far above the threshold. In marked contrast, the figure shows clearly that there is virtually no relationship in the pre-NCLB year between a student's predicted position relative the proficiency threshold and the gain he or she experiences over the predicted score. This is as one would expect, given there was no strong incentive for educators to focus on proficiency prior to NCLB.

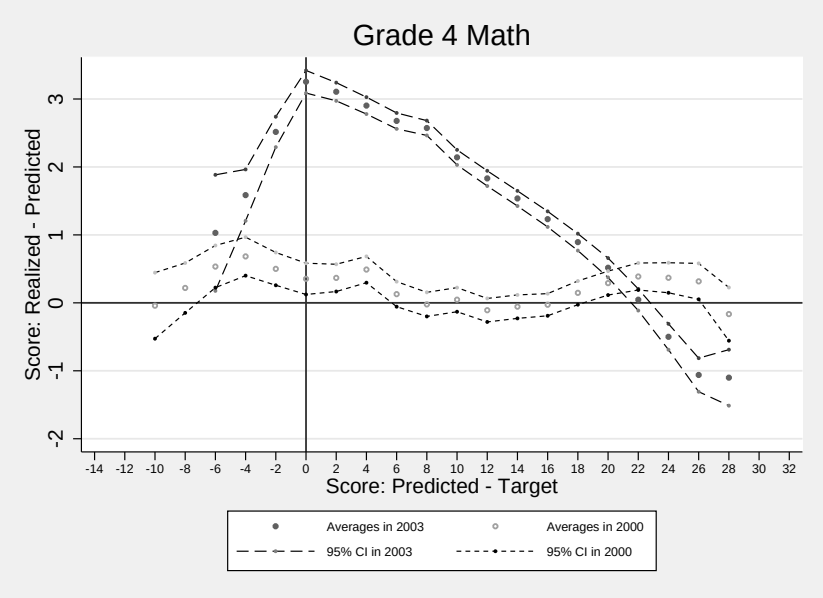
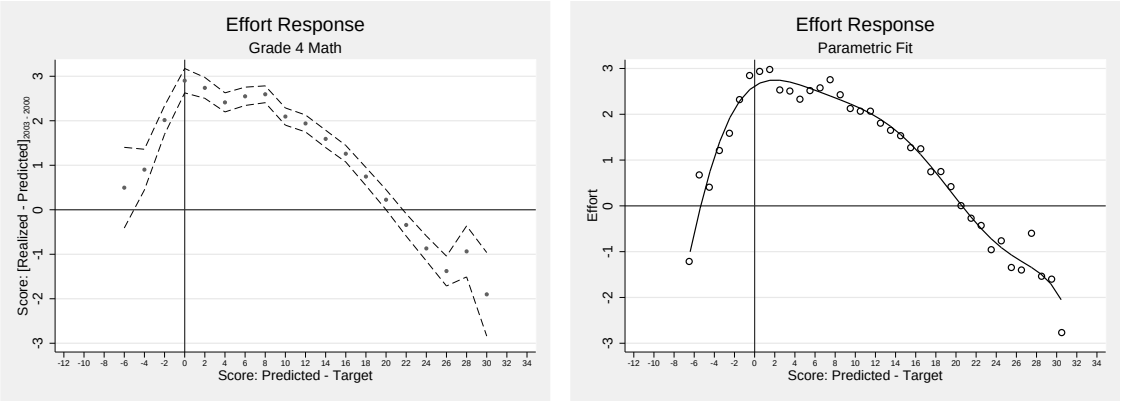


FIGURE 2 – NCLB EFFORT RESPONSE

We are now in a position to recover the effort function. In Figure 3a, we take the difference between the two years – 2003 versus 2000 – to isolate the effort response at each point in the predicted test score distribution (applying the third and fourth steps described above). In Figure 3b, we then fit a flexible polynomial to the data, which we interpret as the effort function, $e^*(\pi)$. We estimate the function by first grouping students in each year into incentive strength bins of width one (in terms of developmental scale units) and calculating the average effort response within each bin. We then take the difference between the 2003 and 2000 averages, weight each difference by the number of students in the bin, and regress the weighted

differences on a flexible eighth-order polynomial of the incentive strength measure, π. The points in Figure 3b represent the within-bin differences and the function is the eighth-order polynomial fit.

The function behaves as theory would predict, peaking where incentives are strongest and steadily declining as incentives weaken.[25] With this function in hand, we can compute the expected effort response for students at any point in the π distribution, under the standard separable production technology assumption used in the literature.



(a) 2003 minus 2000 (b) Parametrized Effort Function

FIGURE 3 – DERIVATION OF THE EFFORT FUNCTION

To the extent that educators care about incentives under the new regime, adjusting discretionary effort is an obvious candidate input through which performance can be altered, and in a manner consistent with the observed change in the test score profile. This 'effort' could take a variety of unobserved forms: teachers raising their energy levels and delivering material more efficiently inside the class-

[25]The asymmetric shape of the effort function is a natural consequence of the first-order condition given in equation (13) given in the Appendix A. The reasoning is best explained using the intuition provided by Figure A.1. Starting from a low θ_L (which is the theoretical analogue to a low $\hat{y}$, and thus low value of the horizontal axis in Figure 3), increasing θ to θ_M shifts the marginal benefit curve to the left and raises effort. However, due to the way in which the marginal benefit and marginal cost curves intersect, one must increase θ a great deal more to obtain the same effort level (with θ_H) as implied by θ_L. Thus, the effort function features a steeper slope on the left-hand side of the target than the right-hand side.

room, increasing their lesson preparation outside the classroom, or teaching more intensively 'to the test.' Without richer data, these various components are difficult to distinguish.

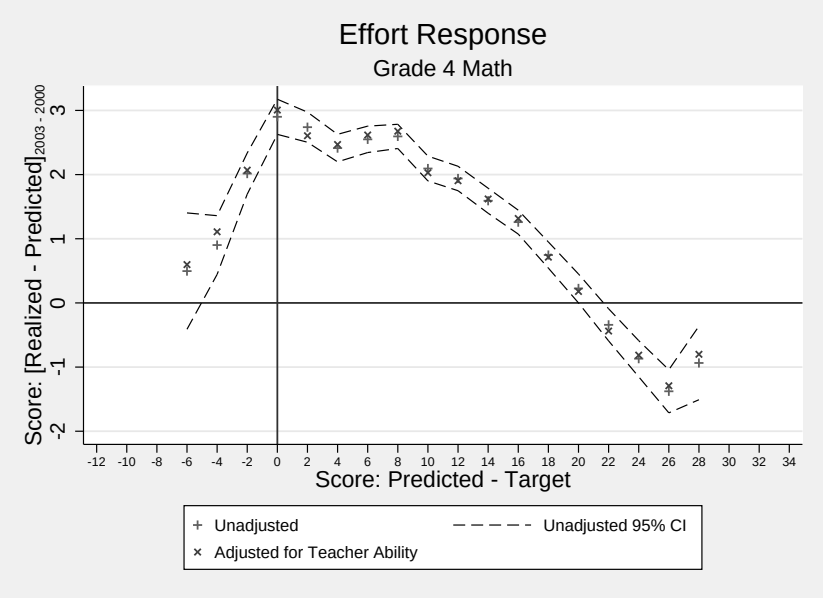
In the next subsection, we argue that the data patterns do not reflect school-level decisions to lower class sizes, differentially group students, or reassign teachers. This leads us to take the evidence as supporting the view that teachers changed their effort in response to NCLB, and in a way according with the hypothesized response to a proficiency-count system.

V.C. Understanding the Effort Response and Ruling out Rival Stories

In this subsection, we show the inverted-U shape in Figure 2 is likely due to educators increasing their *effort* in the vicinity of the passing threshold, rather than adjustments to other relevant inputs to education production. We do so by considering two potentially important alternative channels, involving changes in teacher quality and class characteristics (size and homogeneity) respectively. According to the first, schools sort students differentially to teachers in response to NCLB's introduction based on teacher ability. They might do so if higher-ability teachers were better able to improve marginal students' test scores, providing an incentive to re-assign marginal students to such teachers. This should be visible in our rich longitudinal data. According to the second, schools may sort students differentially to classes based on classroom characteristics, notably class size or classroom homogeneity, in response to NCLB's introduction. They might do so if, for example, schools believed marginal students perform better in small classes or in classrooms with relatively homogenous students. Again, the relevant adjustments should be observable in our data. We take these two sorting channels in turn.

Sorting Based on Teacher Ability. Figure 4 presents evidence relevant to the notion that schools sort students differentially in response to NCLB's introduction to teachers based on ability. We construct the figure as follows: First, we difference

the 2003 and 2000 profiles of raw means in Figure 2 and plot the resulting mean differences along with the associated confidence intervals. We then calculate adjusted means in each incentive strength bin for both 2003 and 2000 by controlling for the effect of teacher ability. We do so by regressing gains above predicted scores on a mutually exclusive and exhaustive set of indicators for the bins on the horizontal axis, fully interacted with year-2003 and year-2000 fixed effects, while controlling additionally for teacher ability fully interacted with year fixed effects[26] Within each bin, we then construct the difference in mean gains across years by subtracting the estimated coefficient on the year 2000 indicator from the estimated coefficient on the year 2003 indicator.



Notes: These figures present the profile showing the difference between the 2003 and 2000 profiles of raw means in Figure 2 and the associated confidence intervals *alongside* the difference between the 2003 and 2000 means adjusted for the effects of teacher ability according to the procedure described in the text.

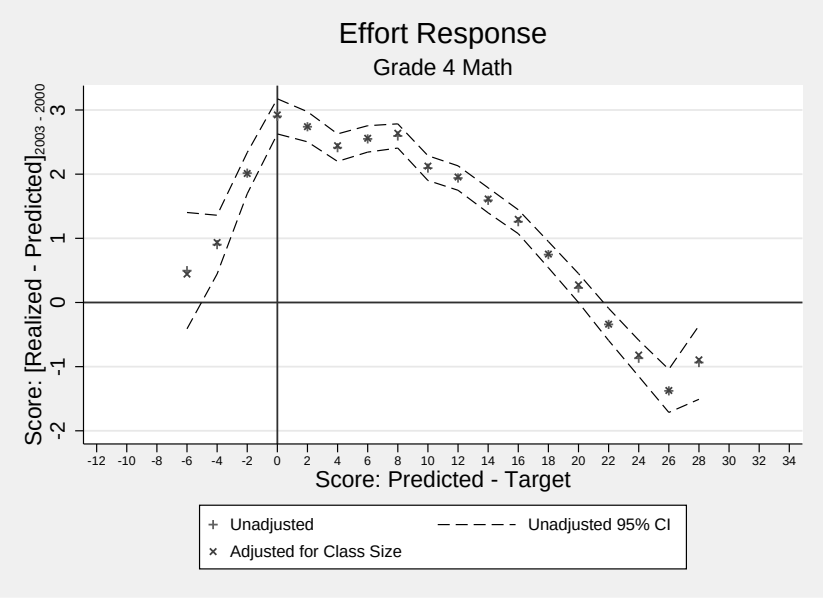
FIGURE 4 – THE IMPACT OF DIFFERENTIAL SORTING BY TEACHER ABILITY

The profile of adjusted mean differences falls entirely within the confidence band for the profile of raw mean differences. This supports the view that teacher

[26]Note that regressing gains above predicted scores in this way but *without* controlling for teacher ability recovers the raw means in each year exactly.

ability does not affect scores *differentially* at any point in the distribution, comparing 2003 and 2000. It is therefore unlikely that schools responded to NCLB by sorting students differentially to teachers based on teacher ability in a way that could explain the test score patterns we observe.

Sorting Based on Class Size. Figure 5 presents evidence relevant to the idea that schools may sort students differentially in response to NCLB's introduction to classrooms based on class size. It does so by repeating the analysis in Figure 4 but adjusting the means in each incentive strength bin by controlling for the effect of class size. We again find that the profile of adjusted mean differences is entirely within the confidence intervals for the profile of raw mean differences, lending strong support to the view that schools did not respond to NLCB by sorting students to classrooms differentially based on class size.



Notes: These figures present the profile showing the difference between the 2003 and 2000 profiles of raw means in Figure 2 and the associated confidence intervals alongside the difference between the 2003 and 2000 means that are adjusted for the effects of class size.

Figure 5 – The Impact of Differential Sorting by Class Size

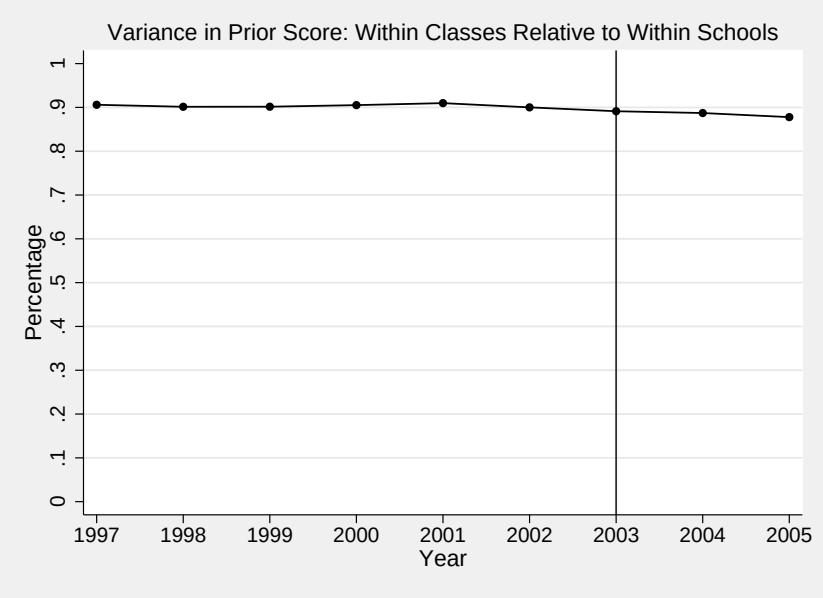
Sorting Based on Lagged Test Scores. We also examine whether schools responded to NCLB by making classrooms more homogeneous. Creating classes in

which students have similar academic preparedness could make it easier for teachers to target instruction toward a particular subset of students without necessarily increasing overall teaching effort. We examine whether students became more similar within classes by examining the relative changes in the within-school and within-classroom variances in prior-year student test scores. If classrooms became more homogeneous, we expect the fraction of within-school variation in that is explained by within-classroom variation to fall in the post-NCLB period.

Figure 6 plots the fraction of the within-school variance in prior-year test scores that is explained by the within-classroom variance in prior-year test scores over time. Overall, the within-classroom variance accounts for approximately 90 percent of the within-school variance, leaving only 10 percent of the within-school variance occurring *across* classrooms. Over time, there is no discernible change to this fraction in 2003, implying that schools did not respond to NCLB by grouping students into more or less homogeneous classrooms. As a point of reference, in 2003, 84 percent of the overall variance in the prior score occurs within schools. Of that, 76 percent of the overall variance occurs within classrooms.

In sum, the observed patterns do not appear to be due to differential sorting of either kind, and are consistent with plausible theories of the way educators adjust effort in response to proficiency-count incentive schemes. This evidence helps justify the model's focus on the effort-setting decision rather than changes in other inputs.

In the Appendix, we also consider an alternative story whereby effort might vary with respect to a student's relative position in the predicted score distribution in his or her school, instead of his or her relative position to the NCLB target. A natural test is to look at the position of peak effort across schools with varying distributions of the predicted score. As we show, the evidence supports the view that schools indeed respond to a student's proximity to the proficiency threshold, and not the relative position of the student in the school-specific predicted score distribution.



Notes: This figure presents, in each year, the ratio of the within-classroom variance of the prior year test score to the within-school variance of the prior year test score.

FIGURE 6 – FRACTION OF WITHIN-SCHOOL VARIANCE IN PRIOR SCORE EXPLAINED BY WITHIN-CLASSROOM VARIANCE

V.D. Anticipating a Structural Model: Motivating Evidence

We take the findings already presented in this section as strong grounds for thinking that educators responded to the introduction of NCLB by increasing effort. The semi-parametric approach we have implemented develops this notion, showing – under minimal assumptions – how effort varies throughout the incentive strength distribution, as captured by the effort function in Figure 3b.

For the remainder of the paper, we wish to explore this effort-incentive strength relationship further. We do so in two parts. First, we write down, then estimate, an explicit model of effort setting (rather than a model of teacher and student classroom assignments), prompted by the evidence just shown. This exercise allows us to determine how effort depends explicitly on primitives of the cost and production technologies as well as incentive scheme parameters. Second, with those elements in hand, we are then able to conduct informative counterfactuals that alter incentives

and allow us to trace the consequences of alternative incentive schemes for the full distribution of score outcomes.

In this subsection, we provide descriptive evidence that motivates us to treat agency at the *teacher* level in the structural model in the next section. That effort decisions are taken at the *teacher* level, even though the incentives under NCLB apply to schools as a whole, is somewhat surprising. A natural interpretation, which we develop at the end of this section, is that school principals delegate 'local' decisions to teachers in service of school-level objectives, given that teachers know each of their students better.

The evidence here is in two parts. First, we demonstrate that a student's position in her classroom-level distribution of incentive strength is an important predictor of our empirical measure of NCLB effort, while her position in the school-level distribution plays little role, consistent with educators making decisions about student effort based on 'local' classroom characteristics. Second, and related, we show that our main reduced-form results remain essentially unchanged when we rely solely on within-classroom variation for identification of the non-parametric patterns we presented above in Figure 2.

Students' Positions in the Classroom Distribution

We first examine whether a student's position in her classroom-level distribution of incentive strength is a better predictor of her test score gains than her position in the school-level distribution, given the school she attends.

Intuitively, one would like to compare two students who are in the same position of their school-level incentive strength distribution, but who occupy different positions in their respective classroom distributions of incentive strength. The test, which we implement below, assesses whether such students experienced differing test score gains. To that end, we first group students into quartiles of the state-level predicted score distribution. We also group students into quartiles of their school-

level distribution of incentive strength, as well as their classroom-level distribution of incentive strength. To compare the importance of classroom and school characteristics in determining the effort students receive, we then restrict the sample to students who occupy different quartiles in their classroom- and school-level distributions and explore which quartile positions predict students' gains over predicted scores.

The actual NCLB target makes the lowest-performing students the most marginal, and Figure 2 shows that our empirical effort measure ("empirical effort") is highest among them. When we regress that measure – a student's observed mathematics scores less his or her predicted mathematics score ($\hat{y}$) – on indicators for quartiles of the state-level distribution of $\hat{y}$ in column (1) of Table 1,[27] we find that effort decreases progressively for students occupying higher positions in the distribution.

Next, we explore whether knowing a student's position in the classroom distribution has any predictive power over and above the position in the state-level distribution. The results in column (2) show that the classroom distribution is important, as empirical effort is decreasing in a student's position in the classroom distribution and the p-value for the test of joint significance of the classroom indicators is approximately zero, strongly rejecting the hypothesis that the position in the classroom distribution is irrelevant.[28] To put the magnitudes in perspective, a student occupying the top quartile of the state-level distribution experiences gains over predicted scores that are 0.32 standard deviations lower than those in the baseline category; a student who also occupies the top quartile of his or her classroom-level distribution experiences gains that are a further 0.15 standard deviations lower.

When we test whether a student's position in the school-level distribution is an important predictor (see column (3)), we find that the school-level indicators are

[27]Students in the first quartile are the omitted baseline category. The average test score gain among them is 2.87 developmental scale points or 40 percent of a standard deviation.

[28]In this case, the omitted baseline group consists of students who are in the first quartile of both the state distribution and their classroom distribution.

Table 1 – The Importance of Student Positions in Classroom- and School-Level Incentive Strength Distributions

Dep. Var.: $y_{it} - \hat{y}_{it}$	(1)	(2)	(3)	(4)
State-Level Distribution				
StateQ2	-0.2958***	-0.2336**	-0.3052***	-0.2527**
	(0.0981)	(0.1010)	(0.1145)	(0.1160)
StateQ3	-1.0317***	-0.8347***	-1.0302***	-0.8935***
	(0.1110)	(0.1236)	(0.1509)	(0.1563)
StateQ4	-1.8971***	-1.6557***	-1.9076***	-1.7496***
	(0.1297)	(0.1485)	(0.1942)	(0.2003)
P-value for Test of Joint Significance of State Indicators	0.00	0.00	0.00	0.00
Classroom-Level Distribution				
ClassQ2		-0.2908***		-0.3946***
		(0.0752)		(0.1390)
ClassQ3		-0.4969***		-0.5189***
		(0.0914)		(0.0981)
ClassQ4		-0.7716***		-0.8731***
		(0.1161)		(0.1607)
P-value for Test of Joint Significance of Classroom Indicators	-	0.00	-	0.00
School-Level Distribution				
SchoolQ2			0.0811	-0.0882
			(0.0991)	(0.1601)
SchoolQ3			0.0015	0.0769
			(0.1292)	(0.1335)
SchoolQ4			0.0703	0.0466
			(0.1668)	(0.2033)
P-value for Test of Joint Significance of School Indicators	-	-	0.55	0.64
Observations	22,199	22,199	22,199	22,199

Notes: The sample is restricted to students who occupy different positions in their classroom- and school-level distributions of incentive strength, respectively. The dependent variable in each column is our empirical measure of effort: a student's observed mathematics scores less his or her predicted mathematics score. Standard errors are clustered at the school level. *** indicates significance at the 1 percent level; ** indicates significance at the 5 percent level.

neither individually nor jointly significant in explaining student test score gains. In column (4), we include indicators capturing a student's position in all three distributions. Here, the classroom-level and state-level distribution indicators remain stable and highly significant, while the school-level indicators offer no predictive power for student test score gains.

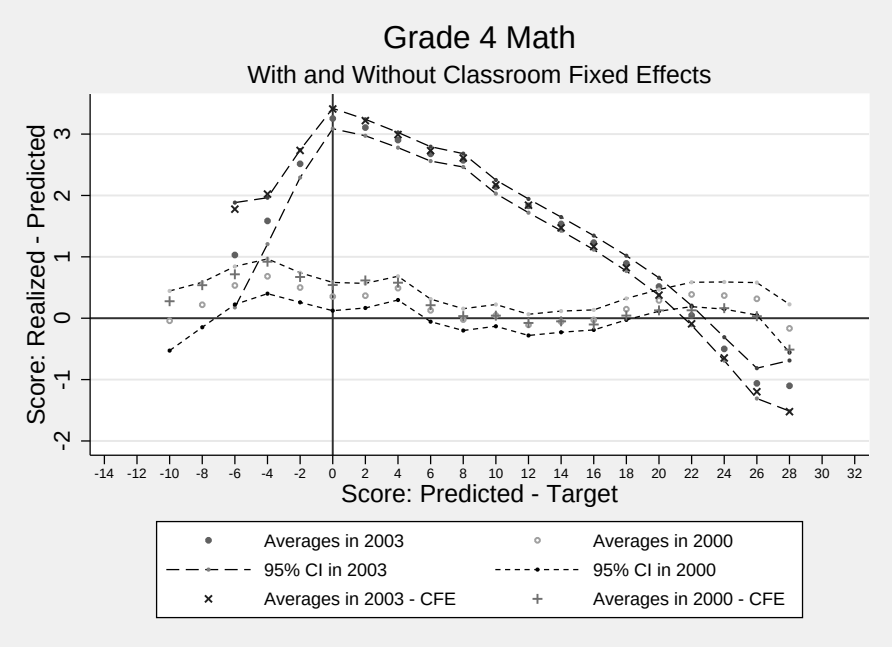
In summary, starting from Figure 2 and trying to improve empirical effort predictions, we see that a student's position in his or her classroom distribution of incentive strength is predictive of gains over predicted scores in a strong, independent way. In contrast, conditional on knowing both a student's position in the state- and classroom-level distributions, a students' position in his or her school-level distribution does not provide further predictive power. Put differently, two students in the same position of their school-level distribution can expect to receive differential effort if they occupy different positions in the distributions of their respective classrooms, as students in the bottom quartile of their classroom-level distributions experience gains that are 0.08, 0.10, and 0.16 standard deviations higher than students in the second, third, and top quartiles, respectively.[29]

Identifying the Non-Parametric Patterns Using Within-Classroom Variation

We further demonstrate the importance of within-classroom variation in incentive strength by showing that our main reduced-form results remain unchanged when we rely solely on within-classroom variation for identification of the non-parametric patterns presented in Figure 2.

Figure 7 plots the raw means in 2003 and 2000 from Figure 2 along with the respective confidence intervals and the adjusted means that prevail after removing across-classroom variation using classroom fixed effects. The adjusted profiles are obtained by regressing (at the student-level) gains above predicted scores on indica-

[29]These results come from column 4 of Table 1 and correspond to 0.40, 0.52, and 0.87 developmental scale points, respectively.



Notes: This figure presents the raw mean test score gains in 2003 and 2000 from Figure 2, along with the confidence intervals. It also depicts the predicted means in each year that are identified using only within-classroom variation in test score gains and incentive strength. To construct the adjusted means in each year using only within-classroom variation, we first regress (at the student-level) gains above predicted scores on a mutually exclusive and exhaustive set of indicators for the bins on the incentive strength axis and classroom fixed effects. Ignoring the estimated classroom fixed effects, we then predict adjusted mean gains as the estimated coefficient on the indicator for each bin.

FIGURE 7 – EMPIRICAL EFFORT USING WITHIN-CLASSROOM VARIATION

tors for the bins plotted on the incentive strength axis *and* classroom fixed effects, and then predicting the mean gains as the estimated coefficient on the indicator for each bin. It is striking that the mean gains predicted using only within-classroom variation lie almost entirely within the confidence intervals of the raw, unadjusted means, supporting the view that the raw data patterns are driven largely by within-classroom variation in incentive strength.

Taken together, the evidence is consistent with a decentralized arrangement in which teachers make effort decisions based on classroom characteristics explains the data well. On that basis, we will develop a model in the next section in which agency is at the teacher level.

It is important to emphasize that we are not taking the position that schools'

administrative teams – the principal, vice principal, etc. – do not respond to NCLB in ways that could be interpreted as effort. Rather, we view the empirical evidence as consonant with a data-generating mechanism in which each individual teacher is told by the school's adminstration which students should receive extra effort and which students should be left alone.[30] That is, it is likely that the entire school, from the top to the bottom (if well-managed), is in agreement regarding how to respond to NCLB, but individual teachers are left to manage their classrooms by determining how best to direct their effort, keeping the agreed-upon overall objective in sight. Given the importance of local classroom conditions in education production, such decentralized effort setting amounts to good management practice.

VI. A Structural Model of Effort Setting

We now present a model of the education process that provides the basis for our structural estimation. For reasons presented in the previous section, we focus on effort setting in the face of accountability incentives, and we treat agency as being at the teacher level. Accordingly, the model links accountability incentives to student performance via teacher effort.

VI.A. Environment

The model has three main elements: an education production technology (including teacher effort as a key input), an accountability incentive scheme, and a cost of effort function. The teacher's objective can be formed from these, allowing us to express the teacher's optimal effort choice as a function of key parameters.

Production Technology. The test score technology relates measured education output y to various inputs. Given our rich data, we focus on the determination of

[30]The opening quotation in Neal and Schanzenbach (2010) encapsulates the type of process perfectly.

individual test scores, denoted y_i for student i.

Among the determinants, we place particular emphasis on the discretionary actions of educators that increase output, given our interest in incentives. In line with a substantial body of work in incentive theory, we will refer to such actions simply as 'effort.'[31] In our setting, we will think of effort as capturing a range of incentive-influenced actions on the part of educators that raise student performance, many of which are unobserved by the researcher. In our empirical implementation, effort will refer specifically to changes in observable test scores attributable to incentive variation rather than changes in other relevant inputs – teacher quality and class size, for instance. We will write the effort directed to student i as e_i, which is endogenous to the prevailing incentive scheme, thus allowing for the possibility that effort within-classroom can be so tailored.

Student scores also depend on various exogenous inputs, such as student ability – we treat students as passive, though potentially heterogeneous. We will summarize these exogenous inputs in a single measure for student i, $\hat{y}_i$. We will think of increases in these inputs as capturing more favorable exogenous 'production' conditions.

Effort and exogenous inputs are assumed to be related in a systematic way to output, measured using test scores. Formally, we write the education production technology as

$$y_i = \hat{y}_i + e_i + \epsilon_i, \tag{3}$$

where $\hat{y}_i$ is the predicted score for student i, e_i is the teacher effort directed to student i, and ϵ_i is a shock to test scores, capturing random determinants unobserved by the econometrician.

[31] The analogy with firms is clear, quoting Laffont and Tirole (1993), page 1: "The firm takes discretionary actions that affect its cost or the quality of its product. The generic label for such discretionary actions is *effort*. It stands for the number of hours put in by a firm's managers or for the intensity of their work. But it should be interpreted more broadly."

We assume that ϵ has a cumulative density function given by $F(\cdot)$, with mean 0 and variance σ^2. In the estimation below, we will assume normality, with $\epsilon \sim N(0, \sigma^2)$. We will also allow for the possibility that teachers' beliefs about the distribution of test score shocks may differ from the true distribution. Specifically, we assume that teachers believe that the distribution of ϵ is normal with mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$. Therefore, teachers believe $\epsilon \sim N(\tilde{\mu}, \tilde{\sigma}^2)$. Let $\tilde{F}(\cdot)$ denote that cumulative density function.

Accountability Incentives. We characterize an *incentive scheme* by a target y^T and a reward b, both exogenously given. (The target, in some instances, may be student-specific, in which case we will append an i subscript.)

This formulation of the target allows for a range of possibilities, considered in more detail below: the target could be an exogenously fixed score, a function of average student characteristics (including past performance), or even be student-specific. The reward parameter b governs how target attainment maps into the educator's payoff, and can include monetary rewards or non-monetary punishments. Specifically, we assume that a teacher receives a benefit b for each student in her class who achieves proficiency status.

Cost of Effort. The teacher teaching classroom c ('teacher c' for short) faces a cost that is convex in effort applied to each student. We assume its functional form is known, given by

$$C(e_1, \ldots, e_{N_c}) = \frac{m}{2}\left[\sum_{i=1}^{N_c} e_i^2 + \theta\left(\sum_{i=1}^{N_c} e_i\right)^2\right]. \tag{4}$$

The m parameter allows the marginal cost of effort to be scaled; we show, however, that it is not separately identified below. The parameter θ governs the extent to which effort choices across students in a given class are independent. In the case where $\theta = 0$, the cost side collapses to one in which education provision conforms

to individualized tuition: estimates of this parameter cast light on whether that extreme case is a useful approximation.

Objective Function. Typically, the objective function for public service providers is difficult to discern, which makes analyzing the behavior of agents working in the public sector difficult; this is in contrast to a firm setting, where profit maximization is often reasonable. An advantage of our application is that an explicit portion of the objective is known. That is a consequence of a formal accountability scheme being in place, the purpose of accountability schemes being to define a clear performance metric, along with explicit rewards and punishments.

We will use that fact to our advantage. Taking the above elements together, we can write down the educator objective under different incentive schemes. We focus on the effort decisions of teacher c, allowing each student to receive student-specific effort, e_i (following the motivating evidence from the previous section).

Teacher c chooses a set of effort levels $\{e_1, \ldots, e_{N_c}\}$ to maximize the following objective, which is the difference between her expected benefit of effort and the effort cost:

$$U = b \sum_{i=1}^{N_c} \Big[1 - \tilde{F}(y^T - \hat{y}_i - e_i)\Big] - \frac{m}{2}\left[\sum_{i=1}^{N_c} e_i^2 + \ \theta\left(\sum_{i=1}^{N_c} e_i\right)^2\right]. \tag{5}$$

Given there are N_c students in the class taught by teacher c, there will be N_c corresponding first-order conditions. Consider the first-order condition for the effort directed toward student i:

$$\tilde{f}\big(y^T - \hat{y}_i - e_i^*\big) = \frac{m}{b}\left[e_i^* + \theta \sum_{j=1}^{N_c} e_j^*\right], \ \forall \ i = 1, \ldots, N_c. \tag{6}$$

Under the normality assumption, the probability density function $\tilde{f}(\cdot)$ is given by

$$\tilde{f}(y^T - \hat{y}_i - e_i) = \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \cdot \exp\left\{-\frac{1}{2\tilde{\sigma}^2} \cdot (y^T - \hat{y}_i - e_i - \tilde{\mu})^2\right\}. \tag{7}$$

to underline the dependence on teachers' beliefs regarding the mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$ of test score shocks, which are not necessarily equal to the true values of these moments.

VI.B. Optimal Effort

For a given test score target, y^T, and predicted score, $\hat{y}_i$, optimal effort for any student i is determined by (6). We will write this as $e^*(\hat{y}_i, y^T, \theta, \widetilde{\sigma}^2, \frac{m}{b}, \widetilde{\mu}, \sigma^2; c)$, making explicit the dependence on the model's parameters ($\theta, \widetilde{\sigma}^2, \frac{m}{b}, \widetilde{\mu}$ and σ^2), and classroom factors – in particular, the classroom-specific distribution of incentive strength, which implies that two students who have the same predicted scores, $\hat{y}$, but are in different classrooms can receive different levels of effort if their classroom-specific distributions of incentive strength differ.

Note that each classroom c has a system of N_c interdependent first-order conditions that must be solved if cost parameter $\theta \neq 0$. That being the case, the marginal cost of effort function implies that the effort applied to one student in the class depends on the effort devoted to all other classmates. The first-order conditions for effort are independent *across* classrooms, though depending on the same parameters $\theta, \widetilde{\sigma}^2, \frac{m}{b}, \widetilde{\mu}$ and σ^2.

In each instance, the structure allows us to express optimal effort as a function of the parameters of the incentive scheme, along with other exogenous characteristics. The model is estimated in a world in which NCLB prevails. We are able to show that the set of classroom-specific optimal effort vectors implied by the model exist and are unique under minimal restrictions on effort that will be satisfied in practice (the effort applied to each student needs to be greater than a threshold

negative level).

VII. Model Estimation and Identification

This section discusses the estimation of the model parameters, followed by the sources of variation that identify those parameters.

VII.A. Estimation

We use equation (12) and the *true* distribution of ϵ_i to estimate the model via maximum likelihood. Recall that the true distribution of ϵ_i is normal with mean 0 and variance σ^2. This, coupled with equation (12), implies that the parameter vector to be estimated is given by $\beta \equiv (\theta, \widetilde{\sigma}^2, \frac{m}{b}, \widetilde{\mu}, \sigma^2)$. For any student i, the individual likelihood function is

$$\begin{aligned} L_i(\beta) &= f(\epsilon_i \mid \theta, \widetilde{\sigma}^2, \frac{m}{b}, \widetilde{\mu}, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{ -\frac{1}{2\sigma^2} \cdot \left(y_i - \hat{y}_i - e^*(\hat{y}_i, y^T, \theta, \widetilde{\sigma}^2, \frac{m}{b}, \widetilde{\mu}, \sigma^2; c) \right)^2 \right\}. \end{aligned} \tag{8}$$

Taking the natural log and summing over all students across the state (N without the 'c' subscript) results in the following log-likelihood function:

$$\begin{aligned} \ell(\beta) &\equiv \sum_{i=1}^{N} \log L_i(\beta) \\ &= -\frac{N}{2} \cdot \log(2\pi) - \frac{N}{2} \cdot \log \sigma^2 - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{N} \left(y_i - \hat{y}_i - e^*(\hat{y}_i, y^T, \theta, \widetilde{\sigma}^2, \frac{m}{b}, \widetilde{\mu}, \sigma^2; c) \right)^2. \end{aligned} \tag{9}$$

The maximum likelihood parameter vector estimate $\hat{\beta}$ is chosen to maximize (9).

We estimate the model using the sample of fourth grade students in 2003 with non-missing test score and predicted test score data and set the proficiency target

equal to the true NCLB target in fourth grade of 247. We also restrict the sample to students in classrooms with at least 7 and no more than 40 students.[32]

VII.B. Identification of Parameters

This subsection discusses the sources of variation that identify the main parameters $\beta \equiv (\theta, \widetilde{\sigma}^2, \frac{m}{b}, \widetilde{\mu}, \sigma^2)$.

For convenience, Table 2 presents a summary of the identifying variation used to estimate the model. We discuss the identification of each parameter in turn, using the table as a guide.

TABLE 2 – PARAMETER IDENTIFICATION

	Parameter				
Source of Identifying Variation	θ	$\tilde{\sigma}^2$	$\frac{m}{b}$	$\tilde{\mu}$	σ^2
Gradient between $\hat{y}$ & Empirical Effort[a]	-	✓	-	-	-
At Least Two Classrooms	✓	-	-	-	-
Within-Class Variance of $\hat{y}$ & Empirical Effort	✓	-	-	-	-
Across-Class Variance of $\hat{y}$ & Empirical Effort	✓	-	-	-	-
Variation in Class Size	✓	-	-	-	-
Maximum of Empirical Effort	-	-	✓	-	-
Location of Empirical Effort Maximum	-	-	-	✓	-
Sum of Squared Deviations from Model	-	-	-	-	✓

Notes: [a] "Empirical Effort" is given by $y - \hat{y}$.

VII.B.1 θ

The parameter θ governs the within-classroom tradeoffs in effort that teachers must make across their students. A positive θ implies that the marginal cost of effort directed to any student i is an increasing function of the effort given to any other

[32]Estimation is done in MATLAB using the 'fmincon' package.

student j. As such, when a teacher decides to direct more effort to student j, it becomes more costly to continue exerting the same levels of effort to all of her other students.

Given that θ dictates these within-class tradeoffs, then trivially there must be more than one student in every classroom (satisfied, given our sample restrictions). There also has to be more than one classroom, as it would otherwise be impossible to separate the effect of a student's absolute position relative to the proficiency threshold from the effect of his or her position relative to the other students the class.

More substantively, identification of θ further relies upon observing both within- and across-classroom variation in incentive strength $\hat{y}$ and 'empirical effort,' $y - \hat{y}$. Within-classroom variation provides the necessary tension across a teacher's incentives to exert effort to each student, as some students are closer to the proficiency target while others are farther way. To then separate the effects of proximity to the proficiency target and a student's position relative to her classroom peers, we require across-classroom variation such that two students with the same absolute value of incentive strength occupy different relative positions in their respective classroom distributions. When there are 'empirical effort' differences across such students, we can infer how the marginal cost of effort varies as a function classroom composition. While not a requirement for identification of θ, variation in class size also contains useful identifying information as, holding all else constant, the marginal cost of effort is higher in larger classes.

VII.B.2 $\tilde{\sigma}^2$

The parameter $\tilde{\sigma}^2$ governs the spread of the effort function or the rate at which effort declines as incentive strength moves away from the level that maximizes effort (in either direction). The maximum likelihood routine selects $\tilde{\sigma}^2$ so that spread of the model's effort function minimizes the distance between empirical effort and

model-implied effort throughout the incentive strength distribution. As such, $\tilde{\sigma}^2$ is identified by the gradient of empirical effort in the direction of incentive strength. Because $\tilde{\sigma}^2$ is the teacher-perceived variance of test score noise, it reflects the uncertainty teachers face when making effort decisions. In general, when uncertainty is fairly low, the gradient of empirical effort is steep: in this case, effort declines quickly in either direction, moving away from the proficiency threshold, because outcomes are predictable and the expected payoff to effort is low. In contrast, when uncertainty is quite high, the gradient of empirical effort is flat, with effort declining slowly in either direction from the proficiency threshold, as the expected payoff to effort is high because of a reasonable probability that below-threshold students will pass and above-threshold students will fail. The maximum likelihood routine adjusts $\tilde{\sigma}^2$ until the gradient of the model's effort function best matches the observed gradient implied by empirical effort and incentive strength.

It is important to note that this discussion does not imply that $\tilde{\sigma}^2$ is pinned down by the covariance of empirical effort and incentive strength. The covariance is a linear operator but the non-parametric relationship we observed between empirical effort and incentive is highly non-linear. This observation necessarily rules out linear regression methods for model identification and calls for the structural estimation routine presented here.

VII.B.3 $\frac{m}{b}$

While $\tilde{\sigma}^2$ governs the model-implied effort function's gradient, the parameter $\frac{m}{b}$ affects its height and is identified, in large part, by the maximum of empirical effort. To see this, notice that one can re-arrange the effort-setting first-order condition by multiplying the probability density function of test score noise by the inverse of $\frac{m}{b}$. Higher values of the parameter $\frac{m}{b}$ therefore correspond to lower peaks for the marginal benefit of effort curve. Since maximal effort is attained at the intersection of the marginal benefit and marginal cost curves at the peak, a low peak implies a

low value for maximum effort while a high peak implies a high value. The maximum likelihood routine selects the parameter $\frac{m}{b}$ so that the resulting model-implied effort profile attains a maximum close to that of the empirical effort profile, thereby minimizing the distance between empirical and model-implied effort.

VII.B.4 $\tilde{\mu}$

The parameter $\tilde{\mu}$ dictates the location of the model-implied effort function's peak. While $\tilde{\sigma}^2$ measures teachers' beliefs over test score uncertainty, $\tilde{\mu}$ represents an alternative way to capture teachers' potential risk aversion. Reflecting the teacher-perceived mean of the distribution of test score noise, one can interpret $\tilde{\mu}$ as a uniform test score shock teachers believe all students will experience. The parameter affords us additional flexibility with which to capture the strong reaction by educators to the introduction of NCLB, helping us pin down which students receive the most additional effort – that is, the location (on the incentive strength axis) of the model's maximum effort level. If teachers behave as if all students will experience a negative shock to test scores, maximum effort is directed toward students who are predicted to score above the threshold, as these students become most marginal in this case. In contrast, if teachers behave as if all students will experience a positive shock, maximum effort is directed toward students who are predicted to score below the threshold. The former case is captured by $\tilde{\mu} < 0$ and the latter by $\tilde{\mu} > 0$. Since the maximum likelihood routine aims to minimize the distance between model-implied and empirical effort, the position of the empirical effort maximum therefore identifies the parameter $\tilde{\mu}$.

VII.B.5 σ^2

One can easily show from the first-order conditions of the maximum likelihood objective function that the parameter σ^2 is equal to the average of the sum of squared deviations between empirical and model-implied effort. As such, the parameter

captures the variability of empirical effort from the model's predictions and, using the production technology, this variance is equivalent to the true variance of test score shocks. Comparing σ^2 to $\tilde{\sigma}^2$ therefore provides a sense for the degree to which teachers over-reacted to NCLB.

VIII. Structural Estimates and Model Fit

Table 3 presents estimates of the model's parameters. The estimates of both $\frac{m}{b}$ and θ are positive, implying that it is costly for teachers to exert effort, and that the marginal cost of effort for any given student is increasing in the amount of effort devoted to other students in the classroom. The estimates also accord with the notion that teachers reacted strongly to NCLB, perhaps even exerting more effort than optimal.

Intuitively, our model captures teachers exerting high effort in response to NCLB's introduction in two ways. First, teachers would have an incentive to try hard if they believed test scores were going to be uniformly lower than normal, which one could think of as shift of the test score shock distribution to the left. That is indeed what we find, as the estimated mean $\tilde{\mu}$ is negative, given by -4.99. Second, teachers could also be led to exert additional effort if they believed test score shocks were more volatile: in this case, to compensate for the potentially large swings in test scores, they would make doubly sure to put students over the threshold by exerting more effort than would be needed if there were lower (perceived) volatility. This is also what we find, as the estimate of $\tilde{\sigma}^2$ is 118 developmental scale points squared, compared to the sum of squared test score deviations from the model – the estimate of σ^2 – being only 15.48 developmental scale points squared.

Embracing the idea that teachers make decisions while considering the characteristics of the students in their classrooms allows us to push this argument further. Because class sizes are usually no more than about 25 students, from the teacher's

Table 3 – Parameter Estimates

Parameter	Estimate
$\frac{m}{b}$	0.0078***
	(0.0003)
$\tilde{\mu}$	-4.9967***
	(0.1503)
$\tilde{\sigma}^2$	118.403***
	(5.9639)
θ	0.0378***
	(0.0029)
σ^2	15.481***
	(0.0661)
N	89,271

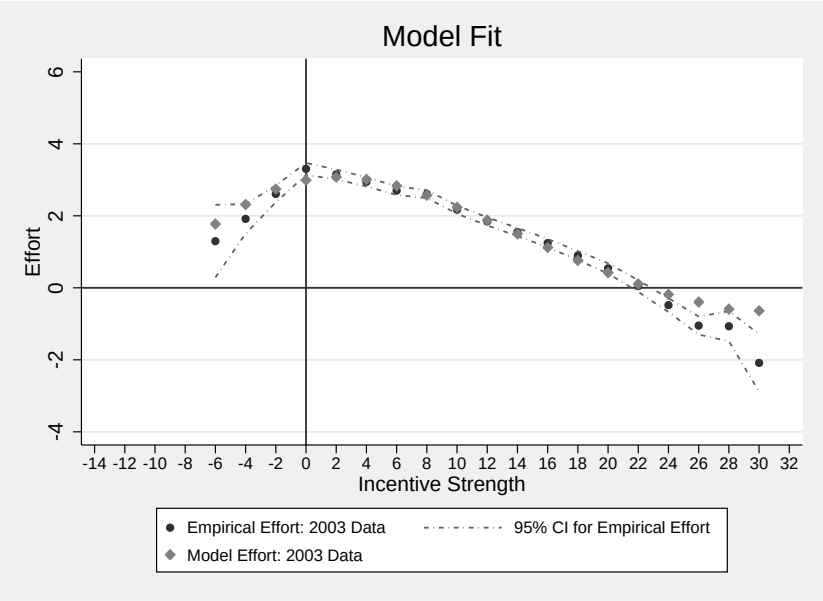
Notes: Standard errors calculated using the Outer-Product of Gradients method appear in parentheses. *** denotes significance at the 1% level;** denotes significance at the 5% level; and * denotes significance at the 10% level.

perspective, cohort-to-cohort variation in the student ability distribution may play an important role in the test-score data generating process. With a relatively small sample, the distribution could experience large changes from year to year and, as a result, in any given year, teachers may actually be unable to form accurate ex-ante beliefs that match ex-post realizations, despite having many years of experience. It is therefore quite natural that we find that teachers appear to be solving a problem with a noise distribution that has a wider variance than the real, state-level noise distribution. The differences in the level of aggregation (classroom vs. state) lead to differences in the role of uncertainty between the teacher's perspective and ours as the econometricians (in the aggregate MLE test score equation).

VIII.A. Model Fit

Figure 8 illustrates model's fit of the data by plotting the year 2003 data from Figure 2 along with the effort predicted by the model. To be clear, the model produces a scatter plot of effort that follows an inverted-U shape. The scatter plot reflects

the idea discussed above that students with the same level of incentive strength can receive different levels of effort depending on the distribution of incentive strength in their classrooms. In Figure 8, we then collapse the scatter plot into binned means along the horizontal axis to aid with the visual presentation. It is clear that the model fits the data very well, as its mean effort prediction is within the confidence intervals of the means from the data in all cases except the far right of the incentive strength distribution.



Notes: This figure presents the profile of empirical effort in 2003 from Figure 2 along with the 95 percent confidence intervals for empirical effort and the binned means of model-implied effort.

FIGURE 8 – INVERTED-U RESPONSE TO NCLB AND MODEL FIT

Table 4 – Model Fit of Proficiency Rates and Test Scores

Subgroup Proficiency Rates and Test Scores	Observed in Data	Predicted by Model
Proficiency Rate		
Overall	0.96	0.96
	(0.19)	(0.19)
White	0.98	0.98
	(0.14)	(0.13)
Black	0.92	0.92
	(0.27)	(0.28)
College-Educated Parents	0.99	0.99
	(0.11)	(0.11)
Non-College-Educated Parents	0.94	0.94
	(0.23)	(0.23)
Economically Disadvantaged	0.93	0.93
	(0.25)	(0.25)
Non-Economically Disadvantaged	0.99	0.99
	(0.12)	(0.12)
Math Score		
Overall	259.51	259.48
	(7.17)	(7.18)
White	261.49	261.49
	(6.82)	(6.84)
Black	255.71	255.73
	(6.36)	(6.32)
College-Educated Parents	262.73	262.73
	(6.70)	(6.79)
Non-College-Educated Parents	257.23	257.23
	(6.61)	(6.55)
Economically Disadvantaged	256.54	256.49
	(6.74)	(6.78)
Non-Economically Disadvantaged	261.99	261.99
	(6.51)	(6.45)

Notes: This table presents observed and model- predicted proficiency rates and test scores for both the overall sample and several sub-samples

Table 4 shows how the model matches various moments of the data. All numbers are rounded to 2 decimal places. The fit is tight, with the model and the data usually differing only at the third decimal place. A further statistic that is not reported above indicates that the within-classroom variance of the realized mathematics score accounts for 74 percent of the overall variance (across all students). Alongside that, the within-classroom variance of the predicted mathematics score from the model accounts for 78 percent of the overall variance, implying that the model replicates the sources of test score variation in the data in an accurate way.

Figures C.1 and C.2 in the Appendix further illustrate the model's fit by plotting the full distributions of observed and model-predicted test scores for both the full sample of students and various sub-samples. As can be seen there, the model fits all test score distributions quite well.

IX. Counterfactual Framework

We are interested in exploring the impact of alternative accountability schemes on the *full* distribution of student outcomes, especially schemes that have yet to be implemented and which may be of policy interest. To that end, we develop a simulation framework based on the structural model and our estimates, using this framework to conduct policy-relevant counterfactual simulations.

In this section, we present our counterfactual framework, then describe how the counterfactuals are comupted. We turn to the counterfactual results themselves in the following section.

IX.A. Framework

Our simulation framework draws heavily on the model, presented in Section VI. It has three elements: an education production technology, an effort-setting condition, and distributional assumptions related to test score shocks.

Technology: We use the same production technology as in (3). Thus, the test score outcome for student i at time t, is given by

$$y_{it} = \hat{y}_{it} + e_{it} + \epsilon_{it}, \tag{10}$$

where $\hat{y}_{it}$ is the student's predicted score, based on all prior information, e_{it} is the effort level directed to that student, and ϵ_{it} is an error term that reflects unobserved determinants of the test score.

Our approach encompasses non-linear technologies, a point we develop below.

Effort: Our primary focus is on the way that teacher effort is influenced by accountability incentives. Thus, we will think of effort as being the result of a teacher optimization problem, written e^*, described in Section VI above. Drawing on the model and estimates, we can determine optimal teacher effort for each student i and indeed, all students currently in that teacher's classroom. This is set according to the effort-setting first-order condition (equation (6)), and is allowed to depend on prevailing incentives, discussed below.

In general, there is no closed-form solution for effort, but equation (# in model) defines teacher effort as a function of the model parameters $\{\theta, \tilde{\sigma}^2, \frac{m}{b}, \tilde{\mu}\}$, one of which relates to the accountability scheme bonus, and a test score proficiency target, y_{it}^T, which may be student-specific.

Test score shocks: We assume that the test score shock faced by student i is given by $\epsilon_i \sim N(0, \widehat{\sigma}^2)$, where we estimate σ^2,

IX.B. Counterfactual Approach

With that basic structure in place, our counterfactual analyses involve setting different student proficiency targets and bonus payments.

The counterfactual analysis will allow us a good deal of flexibility in specifying alternative targets and bonuses. At a general level, we can make targets and bonuses student-specific. Thus we will write the target for student i at time t as y_{it}^T and the bonus, $b_i \equiv b \cdot w_i$, where w_i is the weight for student i: the baseline case with constant bonus payments sets the weight $w_i = 1$ for all students i.

Effort Setting

Having specified the relevant incentive scheme parameters, we use equation (# in model) to explore how effort decisions change.

In period t, teacher c chooses a set of effort levels $\{e_{1t}, \ldots, e_{N_c t}\}$, one for each student in her class, to maximize the following objective, given by the difference between her expected benefit of effort and her effort cost:

$$U = b \sum_{i=1}^{N_c} \left[1 - \tilde{F}(y^T - \hat{y}_{it} - e_{it})\right] - \frac{m}{2}\left[\sum_{i=1}^{N_c} e_{it}^2 + \theta\left(\sum_{i=1}^{N_c} e_{it}\right)^2\right], \tag{11}$$

where N_c denotes the total number of students in teacher c's classroom. The first-order condition for the effort devoted to any student j is given by (6) in the structural model.

For a given test score target, y_{jt}^T, and predicted score, $\hat{y}_{jt}$, optimal effort for student j, given by $e^*\big(\hat{y}_{jt}, y_{jt}^T, \widehat{\mathbf{\Gamma}}; e_{-jt}^*(c)\big)$, is implicitly determined as the solution to equation (6). This effort will depend on both the model's estimated parameters, $\widehat{\mathbf{\Gamma}} \equiv [\frac{\widehat{m}}{b}, \widehat{\tilde{\mu}}, \widehat{\tilde{\sigma}}^2, \widehat{\theta}]$, and the optimal effort levels allocated to all *other* students in the classroom, written $e_{-jt}^*(c)$. When solving for the effort given to any student, we therefore must take into account the classroom-specific distribution of incentive strength, $\hat{y}_{it} - y_{it}^T$, for each student i in the classroom.

Let C denote the total number of classrooms in the data. Determining the full distribution of student effort then involves solving C sets of first-order conditions, with each classroom c having a set of N_c first-order conditions, one for each student. For cost parameter $\theta \neq 0$, the marginal cost of effort for any given student depends on the effort given to all other students in the class. The first-order conditions for effort are independent *across* classrooms (aside from depending on the same parameters, $\frac{m}{b}, \tilde{\mu}, \tilde{\sigma}^2,$ and θ).

For a given set of proficiency targets across students, backing out the distribution of effort is straightforward, once the model parameters have been estimated. We simply solve a system of first-order conditions within each classroom. We do this when computing the effort distribution that prevails under the real NCLB proficiency target and every counterfactual proficiency target we explore. For each set

of proficiency targets we consider, we recompute the effort distribution by solving the system of first-order conditions in each classroom (evaluated at the proficiency targets under consideration) and taking as fixed the distribution of student predict scores and the estimated values of model's parameters.

Counterfactual Output

In turn, using the production technology and the distribution of test score shocks, the test score implied by the model for any student i under proficiency test score target y_{it}^T and bonus payment regime $b \cdot w_i$ is

$$y_{it} = \hat{y}_{it} + e^*(\hat{y}_{it}, y_{it}^T, \widehat{\mathbf{\Gamma}}; c) + \epsilon_i, \tag{12}$$

where $\widehat{\mathbf{\Gamma}} \equiv [\frac{1}{w_i}\widehat{(\frac{m}{b})}, \widehat{\tilde{\mu}}, \widehat{\tilde{\sigma}}^2, \widehat{\theta}]$ and $\epsilon_i \sim N(0, \widehat{\sigma}^2)$. The '$c$' after the semi-colon in the effort function indicates that effort depends on the distribution of incentive strength within student i's classroom, as indicated in equation (6).

Equation (12) can be used to recover the associated test score distribution across all students. Although we take the characteristics of the student population as given throughout our simulations – for example, not modelling inflows from or outflows to private schools in response to NCLB – the dependence of effort on classroom characteristics allows us to incorporate the effects of potentially changing student characteristics, both within and across classrooms.

Types of Scheme

The counterfactual analyses we carry out involve a variety of alternative incentive schemes, which we describe in some detail.

At a general level, accountability schemes can be characterized by a set of targets and rewards, given prior information I. In our application, we treat the predicted score ($\hat{y}$) using all prior information available to the econometrician as

our summary measure of I, thus writing the general class as $\{y^T(\hat{y}), b(\hat{y})\}$ – the schemes used most widely in practice (discussed next) are special cases.

Fixed schemes: These involve targets that are the same for all students – for example, those in a certain grade, as is the case under NCLB. Let the proficiency target that applies under a fixed scheme be y^T. The fixed scheme pays out according to a threshold rule, given by $b \cdot 1(\, y_{it} \;\geq\; y^T)$, where b is the reward if student i's test score at time t, y_{it}, exceeds the student-invariant target y^T (or the sanction if the score does not exceed the target, as under NCLB).

In terms of the fixed target counterfactuals we consider, the actual NCLB target provides a useful benchmark: we explore the effects of setting targets that are higher or lower in the predicted score distribution than this actual target, using our model to determine the associated effort decisions and implied test score distribution in each counterfactual instance.[33]

Student-Specific Bonus Payments: Within the class of fixed schemes, our framework allows us to consider counterfactual regimes that make *student-specific* bonus payments, unlike any scheme currently in operation. Allowing for student-specific bonus payments $b \cdot w_i$, as described above, affords policymakers an additional degree of freedom with which to improve outcomes. We consider two highly contrasting cases: In the first, higher weight (in the form of a higher student bonus) attaches to lower-performing students, with the weight decreasing linearly in students' predicted scores.[34] In the second case, the weight increases linearly in students' predicted scores, creating incentives to favor higher-performing students.[35]

(In the Simulation Appendix, we set out the rationale for the chosen bonus-payment functional forms, and explain how we modify the cost-equating procedure

[33] See Simulation Appendix for details.

[34] In this case, we treat the student-specific bonus payment as $b(\hat{y}_i) = b\frac{(\hat{y}_{\max}+1-\hat{y}_i)}{\hat{y}_{\max}-\hat{y}_{\text{med}}+1}$, where $\hat{y}_{\max}$ is the maximum value of $\hat{y}_i$ across all students in the state and and $\hat{y}_{\text{med}}$ is the median value of $\hat{y}_i$.

[35] In this case, we model the student-specific bonus payment as $b(\hat{y}_i) = b\frac{(\hat{y}_i-y_{\min}+1)}{\hat{y}_{\text{med}}-\hat{y}_{\min}+1}$, where $\hat{y}_{\min}$ is the minimum value of $\hat{y}_i$ across all students in the state.

in these two cases.)

Value-added (VA) schemes: These set targets that are student-specific, depending on a student's prior-year test score, $y_{i,t-1}$. The VA target for student i is given by $y_{it}^T = \delta + \alpha y_{i,t-1}$, and the relevant threshold benefit rule can be written $b \cdot 1(\, y_{it} \geq y_{it}^T)$. The parameter δ determines the mean of the incentive strength – or $(\hat{y}_{it} - y_{it}^T)$ – distribution, while α governs the variance of that distribution; b is again the reward if the test score y_{it} exceeds the target, or the sanction otherwise.

We explore the effects of different VA targets on outcomes by varying the target parameters systematically, as follows: Each value-added target we consider is linked in a precise way to a corresponding fixed target, a fact that facilities comparisons between fixed and value-added schemes in a way that is informative for policy. Here, we make use of the fact that a fixed target is a special case of a VA target, where $\alpha = 0$ and $\delta = y^T$. Thus, taking a given fixed target (the NCLB benchmark of 247, for example), then for any multiplicative coefficient, α, we choose $\delta(\alpha)$ such that the mean of the resulting incentive strength distribution under the VA target matches the mean under the given fixed target.[36]

In the counterfactuals below, for each fixed target we analyze, we consider twelve different settings of the multiplicative coefficient, α. In doing so, we place more or less emphasis on the prior score.[37] For each α under consideration, we then choose the corresponding level of δ to ensure the mean of the resulting incentive strength distribution is equivalent to the mean under the associated fixed target, the (α, δ) pairs giving the twelve different VA targets we examine for each fixed target. (The Simulation Appendix provides more detail regarding the fixed and VA

[36]Taking the NCLB benchmark, for instance, we have $\delta = 247 - \alpha \bar{y}_{t-1}$, implying that the mean of the VA targets (across all students) is also 247 and the mean of incentive strength under the VA target matches the mean under the fixed target.

[37]Specifically, we consider multiplicative coefficients in the range 0.1 to 1.9 – see Simulation Appendix.

targets we analyze.)

Cost equivalence

For comparability, we wish to place all the counterfactual incentive schemes we consider on a common footing. To that end, we make sure that every target regime results in the same cost[38] by changing the bonus payment b until we achieve cost-equivalence across regimes.[39]

We describe the essence of the cost-equating procedure here – a fuller description is available in the Simulation Appendix.

Teachers' optimal effort choices are influenced by the parameter $(\frac{m}{b})$, which appears in the effort-setting first-order condition. While we cannot separately identify the bonus payment b in our estimation framework, we normalize it to one and effectively change it by multiplying the estimate of $(\frac{m}{b})$ by a constant $\frac{1}{k}$: setting $k < 1$ is equivalent to decreasing the bonus payment and setting $k > 1$ is equivalent to increasing the bonus payment. Under each target regime, we pick the value of k that equates the cost to the cost prevailing under the actual NCLB target.

Having ensured cost-equivalence across regimes, we then compare the effort decisions and test score outcomes that result from alternative fixed and value-added targets.

X. Counterfactual Results

In this section, we present the results of our counterfactual analyses.

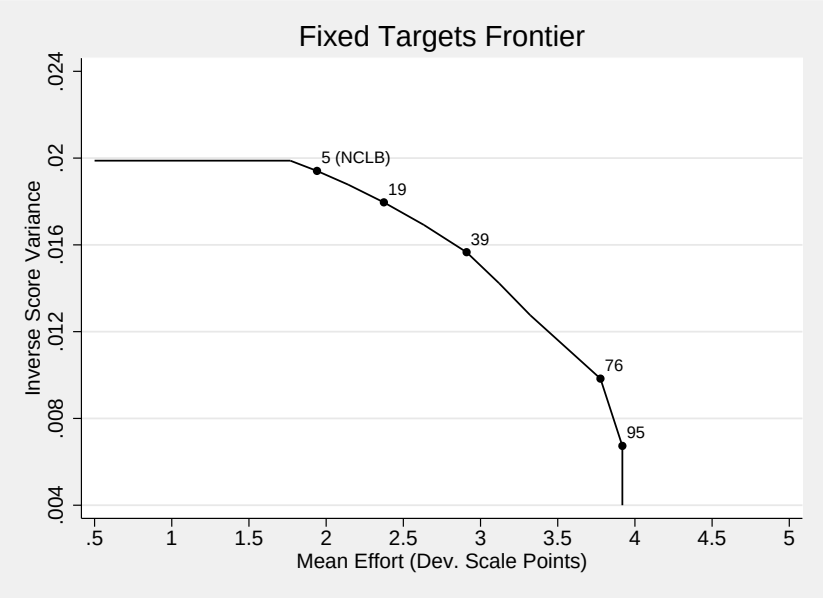
[38]Under a constant bonus scheme, this is equivalent to a given statewide proficiency rate, recalling that the state must pay a bonus for each student who is deemed proficient.

[39]The target cost that all regimes are equated to involves a proficiency rate of 0.96, the observed rate in 2003 under the actual NCLB target.

X.A. Fixed Targets with Homogeneous Bonus Payments

Here we show that fixed targets give rise to a tradeoff between average teacher effort and test score inequality: targets that allow higher effort to be attained come at the expense of greater test score inequality.

This in a new result in the education literature. We demonstrate it by using our model to trace out a frontier in the space of mean effort and a measure of spread – the inverse of the test score variance.[40] Changing the proficiency target from the real NCLB target to other percentiles in the predicted score distribution and plotting the resulting 'mean effort-inverse variance' points traces out the frontier in Figure 9.



Notes: Each point on the frontier reflects the mean effort and inverse test score variance that prevails under a given fixed target (labelled by the percentile of the fixed target in the distribution of student predicted scores). These are calculated by using the counterfactual framework to determine effort decisions and the resulting test score distribution under each fixed target.

FIGURE 9 – FIXED FRONTIER WITH HOMOGENEOUS BONUS PAYMENT

The point associated with the target at the fifth percentile is the only point corresponding to the observed test score variance, reflecting outcomes under the

[40] Taking the inverse implies our inequality measure increases when the outcome is better – in this case, inequality is lower.

actual NCLB proficiency target. All other points on the frontier are simulated using our model. We label five points associated with five separate fixed targets in Figure 9, where target labels correspond to target percentile positions in the predicted score distribution. Table D.1 in the Simulation Appendix shows how the percentile targets reported in Figure 9 map into developmental scale points, the units used by NCLB.[41] The frontier shows a clear tradeoff: higher fixed targets lead to higher mean effort but at the cost of higher test score inequality (or lower inverse test score variance).

The magnitudes involved are quantitatively significant: moving the proficiency target from the 50th to the 75th percentile of the predicted score distribution increases mean effort by 0.09 standard deviations of the test score but at the cost of increasing the test score variance by 45 percent. Setting progressively higher fixed targets is associated with a progressively steeper tradeoff – for example, increasing the target again from the 75th to the 95th percentile increases mean effort by only 0.02 standard deviations but raises the test score variance by 46 percent.

The magnitude of the tradeoff is governed by two parameters: θ and $\tilde{\sigma}^2$. In terms of θ, the slope of the frontier is steeper for higher values of θ, as demonstrated in panel (a) of Table 5. In column (1), we summarize the slopes of the frontiers that result from three different values of θ (low, estimated, and high) with the slopes of the straight lines connecting the mean-effort-inverse-variance points corresponding to targets at the fifth percentile and the median of the predicted score distribution.[42] To put in perspective the magnitudes of the slope changes, suppose one starts at the point associated with the real NCLB target, where the test score variance is 52 developmental scale points squared and mean effort is 1.94 developmental scale points. Increasing mean effort by 1 developmental scale point (14 percent of a

[41] While only these five points are labelled for expositional clarity, we use more fixed targets to trace out the frontier.

[42] Using the slope of the straight line that connects these two specific points is for ease of exposition. Similar results follow from using any other two points to infer the slope along different points of the frontier.

standard deviation) is associated with an increase in test score variance of 9.6, 14.5, and 18.3 developmental scale points squared, respectively, for the low, estimated, and high values of θ.

To understand how the parameter θ affects the slope of the frontier, recall that it governs how the marginal cost of effort for any given student changes with the effort devoted to all other students in the class. As shown in column (2), a high θ therefore leads to greater effort disparities (variance) across students, as it is more costly for teachers to devote effort to any students without relatively strong incentives stemming from the test score proficiency target.[43] High values of θ effectively tighten the marginal window around the proficiency target in the predicted score distribution, implying that progressively higher targets result in the largest test score gains among progressively better students. We demonstrate this point using the target at the median of the predicted score distribution and contrasting between students who are close to the target – those between the 25th and 75th percentile of the distribution – and students how are well below the target and predicted to fail – those below the 25th percentile. Columns (3) and (4) in panel (a) in show that both sets of students receive less mean effort as θ rises but that the gap between the two groups grows, with students who are close to the target receiving relatively more effort for higher values of θ. Similar patterns occur at higher targets, where marginal students are positioned even higher in the predicted score distribution, implying that mean effort gains are accompanied by progressively higher test score variances, leading to exacerbated performance disparities.

In contrast to the the parameter θ, panel (b) of Table 5 shows that the slope of the frontier is *decreasing* in the parameter $\tilde{\sigma}^2$. Starting again at the point associated with the real NCLB target, increasing mean effort by 1 developmental scale point comes at the cost of increasing the test score variance by 34.7, 14.5, and 10.4

[43] Again for expositional clarity, we report the effort variance at one fixed target: the one placed at the median of the predicted score distribution. The same results across frontiers with different θ's hold at other fixed targets.

TABLE 5 – THE ROLES OF θ AND $\tilde{\sigma}^2$ IN EXPLAINING THE SLOPE OF THE FRONTIER

	(1)	(2)	(3)	(4)
	Frontier Slope	Fixed Target at the Median of $\hat{y}$ Distribution		
	(Between 5th and 50th Percentile Targets)	Effort Variance (Dev. Scale Pts^2)	Mean Effort for Students Between Percentiles 25 and 75	Mean Effort for Students Below 25th Percentile
Panel (a): Varying θ				
$\theta_L = 0.5 * \hat{\theta}$	-.0033	1.32	4.76	2.84
$\hat{\theta}$	-.0042	1.62	4.09	1.84
$\theta_H = 1.5 * \hat{\theta}$	-.0055	1.74	3.51	1.10
Panel (b): Varying $\tilde{\sigma}^2$				
$\tilde{\sigma}^2_L = 0.5 * \widehat{\tilde{\sigma}^2}$	-.0077	6.65	5.06	0.69
$\widehat{\tilde{\sigma}^2}$	-.0042	1.62	4.09	1.84
$\tilde{\sigma}^2_H = 1.5 * \widehat{\tilde{\sigma}^2}$	-.0032	0.59	3.23	1.94

In column (1), we report the slope of the frontier that results when we manipulate either $\hat{\theta}$ or $\widehat{\tilde{\sigma}^2}$. Each slope corresponds to the straight line connecting the point corresponding to the real NCLB target and the point corresponding the target at the median of the predicted score distribution. In columns (2), (3), and (4), we hold the target fixed at the median of the predicted score distribution. Column (2) reports the effort variance that results at each value of $\hat{\theta}$ or $\widehat{\tilde{\sigma}^2}$, while column (3) reports the mean effort among students between 25th and 75th percentiles and column (3) reports mean effort among students below the 25th percentile.

developmental scale points squared, respectively, for the low, estimated, and high values of $\tilde{\sigma}^2$.

To see how the parameter $\tilde{\sigma}^2$ determines the slope of the frontier, note that $\tilde{\sigma}^2$ governs test score uncertainty and is identified using the spread of the non-parametric effort profile in the data. As such, it dictates the rate at which effort dissipates from marginal students to those who are predicted to score far below or above the proficiency target, helping determine the effort variance across students. As shown in column (2), when there is little uncertainty – captured by a low value of $\tilde{\sigma}^2$ – the effort variance is high because teachers focus almost exclusively on marginal students, resulting in a steep decline of effort for students in either direction away from proficiency target in the predicted score distribution. In contrast, when there is much uncertainty – captured by a high value of $\tilde{\sigma}^2$ – the effort variance is low because teachers devote similar levels of effort to all students, as the potential payoff to effort for non-marginal students is higher. These patterns, crucial to the slope of the mean-effort-inverse-variance frontier, are demonstrated further in columns (3) and (4), which show that marginal students receive less effort and non-marginal students receive more as $\tilde{\sigma}^2$ increases, owing to the flatter slope of the effort function with respect to the predicted score. Increasing mean effort by raising the target is therefore accompanied by smaller variance increases (inverse variance decrease) at higher values $\tilde{\sigma}^2$ because the relatively flat slope of the effort function ensures that effort gains are relatively less concentrated among (high-performing) marginal students than in cases with low values of $\tilde{\sigma}^2$.

This effort-variance tradeoff is a new finding in the education literature. Prior work has emphasized that proficiency schemes create incentives to focus on students who are marginal with respect to the target, regardless of where in the student distribution the target is set.[44] Marginal students constitute only a fraction of the

[44]See, for example, Reback (2007), Neal and Schanzenbach (2010), Ladd and Lauen (2010), and Deming et al. (2013).

student population, however, and most prior work takes their position in the distribution as fixed (by the accountability scheme in operation). In contrast, our structural approach allows us to explore a host of different proficiency targets and their associated effects on the *entire* distributions of effort and test scores, uncovering the inherent tradeoff present in the choice of targets.

Holding the model's parameters fixed at their estimated values, the effort-variance tradeoff is driven by two forces. The first force – the 'distributional effect' (DE) – captures teachers responses to the target being set at a higher point in the predicted score distribution. The second force – labelled the 'cost-equating effect' (CEE) – reflects how outcomes change when we adjust the bonus payment to equate costs across all target regimes. The shape of the frontier is driven by both forces but in ways that depend on the range in which the proficiency target falls.

When the proficiency target is below the median of the predicted score distribution, the shape of the frontier arises solely from the DE. The CEE also operates in this target range but it is not required for generating simultaneous mean effort increases and inverse variance decreases (the DE does so on its own). In contrast, when the target is above the median of the predicted score distribution the CEE *is* need to generate simultaneous mean effort increases and inverse variance decreases, as the DE alone leads to a decrease in mean effort.

To illustrate these points, Table 6 shows the precise magnitudes of the DE and CEE (on both mean effort and inverse variance) for several fixed targets. Increasing the target while it still below the median makes a progressively larger mass of students marginal, creating sharper incentives for more students.[45] In these cases, the DE to leads to higher mean effort. But increasing the target also makes progressively *better* students marginal, implying that low-performing students receive

[45]If we define a student as 'marginal' if his or her predicted score is within 4 developmental scale points of the target, the fractions of marginal students at targets at the 5th, 19th, and 39th percentiles of the predicted score distribution are 0.19, 0.34, and 0.4, respectively. (Other 'marginal' windows lead to similar patterns.)

relatively little effort, exacerbating performance inequality and increasing test score variance (so decreasing inverse variance). Setting progressively higher targets in this range therefore increases mean effort and raises test score variance through the DE, resulting in the downward-sloping frontier shape depicted in Figure 9.

Table 6 – Decomposition of the Distributional and Cost-Equating Effects in Moving Along the Frontier

	(1)	(2)	(3)	(4)
Target	Mean Effort		Inverse of the Test Score Variance	
(Percentile Position)	Distributional Effect	Cost-Equating Effect	Distributional Effect	Cost-Equating Effect
19	0.24	0.20	-0.0017	0.0003
39	0.27	0.70	-0.0039	0.0002
50	0.21	0.97	-0.0050	-0.0002
76	-0.23	2.06	-0.0071	-0.0025
95	-0.83	2.80	-0.0075	-0.0052

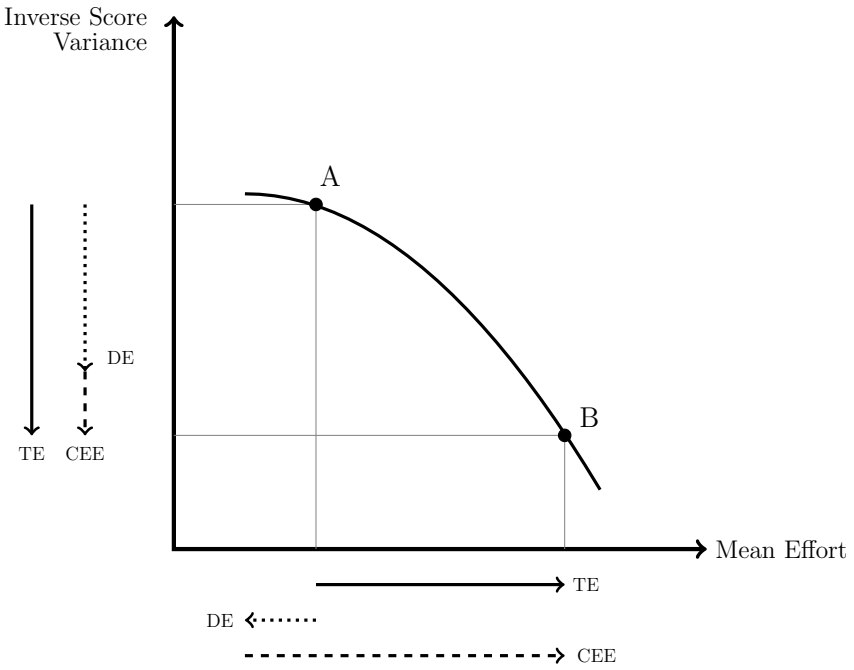
In columns (1) and (2), we present the DE and the CEE on mean effort, respectively, that occur when we move along the frontier in Figure 9 from the point corresponding the the real NCLB target to points corresponding to the other targets on the frontier. In columns (3) and (4), we do the same but report the DE and CEE on the inverse of the test score variance.

When the proficiency target is above the median in the predicted score distribution, raising the target further makes a progressively smaller mass of students marginal, with a progressively larger mass of students being predicted to miss the proficiency target.[46] As shown in Table 6, for higher targets in this range, the DE leads to reductions in mean effort because the targets make it prohibitively costly for teachers to help their students meet proficiency standards.[47] In these cases, the shape of the frontier relies on cost equating across regimes – a point we now elaborate on with the aid of Figure 10.

To explain the importance of the CEE, consider the illustrative Figure 10. Start at point A, which corresponds to the mean effort and inverse test score vari-

[46] Again defining a student as 'marginal' if his or her predicted score is within 4 developmental scale points of the target, the fractions of marginal students at targets at the 76th and 95th percentiles of the predicted score distribution are 0.31 and 0.15, respectively. The fractions of non-marginal students who are predicted to fail are 0.50 and 0.84, respectively.

[47] DE reflects teachers being discouraged by the overly-ambitious standards and responding optimally by exerting less effort (because the probability of target attainment is very low).



Notes: This figure demonstrates the forces that give rise to the shape of the frontier in Figure 9, focusing on the move from a fixed target at the 5th percentile of the predicted score distribution (point A) to a fixed target at the 95th percentile of the predicted score distribution (point B). The label "DE" describes the *distributional effect* that arises from changing the proficiency target and therefore each student's position relative to the target in the predicted score distribution. The label "CEE" describes the *cost-equating effect*, which is the amount that mean effort and the inverse test score variance change as a result of altering the bonus payment to achieve cost-equivalence across target regimes. The label "TE" describes the *total effect*, calculated as the sum of the distributional effect and the cost-equating effect.

FIGURE 10 – INTUITION FOR FIXED FRONTIER WITH HOMOGENEOUS BONUS PAYMENT

ance under the real NCLB target (set at 247 developmental scale points, which corresponds to only the fifth percentile in the distribution of the predicted score). The NCLB target being set low in the predicted score distribution makes relatively low-performing students the most marginal, implying that they receive more effort than high-performing students. Because low-performing students receive a higher boost than high-performing students, the effort disparity results in a low test score variance (high inverse variance), leading to small test score gaps. Suppose that we increase the proficiency target considerably above the median, placing it at (say) the ninety-fifth percentile. Doing so yields a resulting 'mean effort-inverse variance' given by point B.[48] In this case, the DE and the CEE work in different ways.

[48]The corresponding developmental scale point value for the target at the ninety-fifth percentile

Take the DE first: At the higher target, the newly-marginal students will be high up in the predicted score distribution, and it is these students who will receive the most effort. Because other students receive relatively little effort, average effort falls relative to the regime using the real NCLB target (by 0.83 scale points, as indicated in column (1) of Table 6). At the same time, the test score variance increases (so inverse variance falls) because high-performing students receive the most effort, leading to a widening of performance disparities. Thus, if we allowed only the DE to operate without maintaining cost equivalency, the point at the higher fixed target would have lower mean effort and higher test score variance.

Now considering the need to maintain cost equivalency, the higher proficiency target associated with a point like B results in less effort, a lower proficiency rate and correspondingly lower costs. The bonus payment b must therefore be raised to increase effort and equate costs with the benchmark regime associated with the real NCLB target. Cost equating increases mean effort by 2.80 scale points (see column (2) of Table 6) but it further increases the test score variance (decreases the inverse variance), as high-performing students benefit disproportionately from the higher bonus payment, owing to their marginal position in the incentive strength distribution. In this case, cost equating compensates for teachers exerting less effort in response to the target being set too high, pushing mean effort at point B to a level much higher than the mean effort observed at point A and giving rise to the shape of the frontier.

X.B. Fixed Targets with Heterogeneous Bonus Payments

In this subsection, we construct frontiers for the two heterogeneous bonus-payment regimes described above, demonstrating the magnitudes by which outcomes can be improved by redistributing bonus payment money.[49] We do so by placing the

is 269.

[49]Recall: In the first, higher weight (in the form of a higher student bonus) attaches to lower-performing students, the weight policymakers place on students *decreasing* linearly in students'

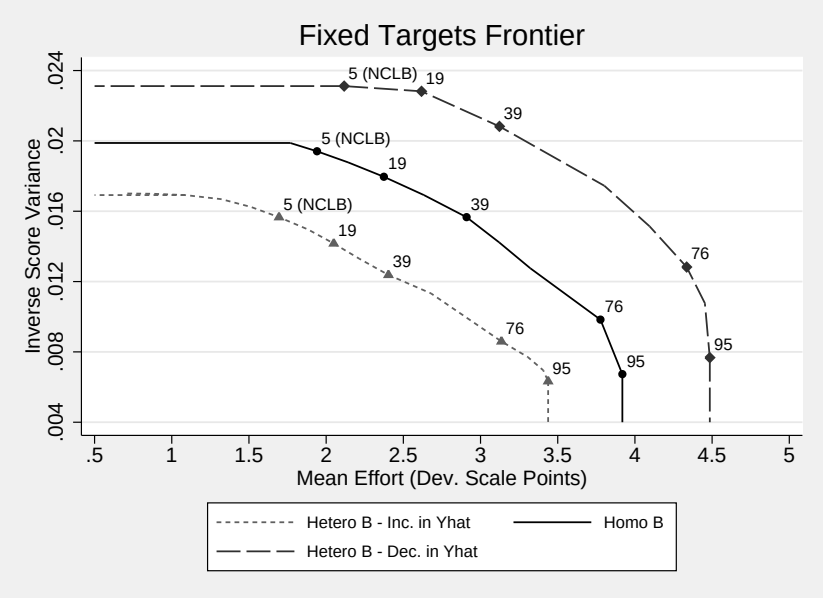
two resulting frontiers alongside the frontier based on homogenous bonus payments (obtained in the previous subsection).

Figure 11 shows that the scheme that places more weight on low-performing students dominates the homogenous bonus payment regime, which in turn dominates the scheme that places more weight on high-performing students. In terms of the magnitudes, holding the proficiency target fixed at the real NCLB value and attaching more weight to low-performing students increases mean effort by 2.5 percent of a standard deviation and decreases test score variance by 16 percent; attaching more weight to high-performing students decreases mean effort by 3.4 percent of a standard deviation and increases test score variance by 24 percent. Averaging across all fixed targets, the corresponding changes are 4.5 percent of a standard deviation more effort and 17.6 percent less variance under the first regime, and 5.3 percent of a standard deviation less effort and 19.3 percent more variance under the second regime.

We explain why outcomes improve when we switch to a regime that places more weight on low-performing students by decomposing the forces that cause the frontier to shift out. Holding the proficiency target fixed at given level, two forces drive the transition from the homogenous bonus payment frontier to the frontier in which the bonus payment is decreasing in the predicted score: the 'bonus payment effect' (BPE), constituting the change that results from switching the bonus payment structure but not ensuring cost equivalence; and the CEE, representing the subsequent change that results from equating costs.[50] The precise magnitude of each force at several different fixed targets is shown in Table 7, while Figure 12 presents a qualitative illustration of the two forces at the real NCLB target (the transition from point A to point A') and the target set at the 95th percentile (the

predicted scores. In the second case, we make the weight *increase* linearly in students' predicted scores, creating incentives to favor higher-performing students.

[50]There is no DE here because the description is geared toward explaining *shifts* of the frontier, *not movements along* a frontier. The DE only comes into play when we consider changing the target and moving to a different point along the same frontier.



Notes: In this figure, each point reflects the mean effort and inverse test score variance that prevails under a given fixed target. The solid line reproduces the homogenous bonus payment frontier shown in Figure 9. The long-dash line shows the frontier that arises when bonus payments are student-specific and linearly decreasing in the predicted score. The short-dash line depicts the frontier that arises when bonus payments are student-specific and linearly increasing in the predicted score. All points are calculated by using our model under the appropriate bonus payment regime to determine effort decisions and the resulting test score distribution for a given fixed target. The point labels correspond to the percentile position of the fixed target in the distribution of student predicted scores.

FIGURE 11 – FIXED FRONTIERS WITH HETEROGENEOUS BONUS PAYMENTS

Table 7 – Decomposition of the Bonus Payment and Cost-Equating Effects in Switching to Bonus Payments that are Decreasing in Predicted Scores

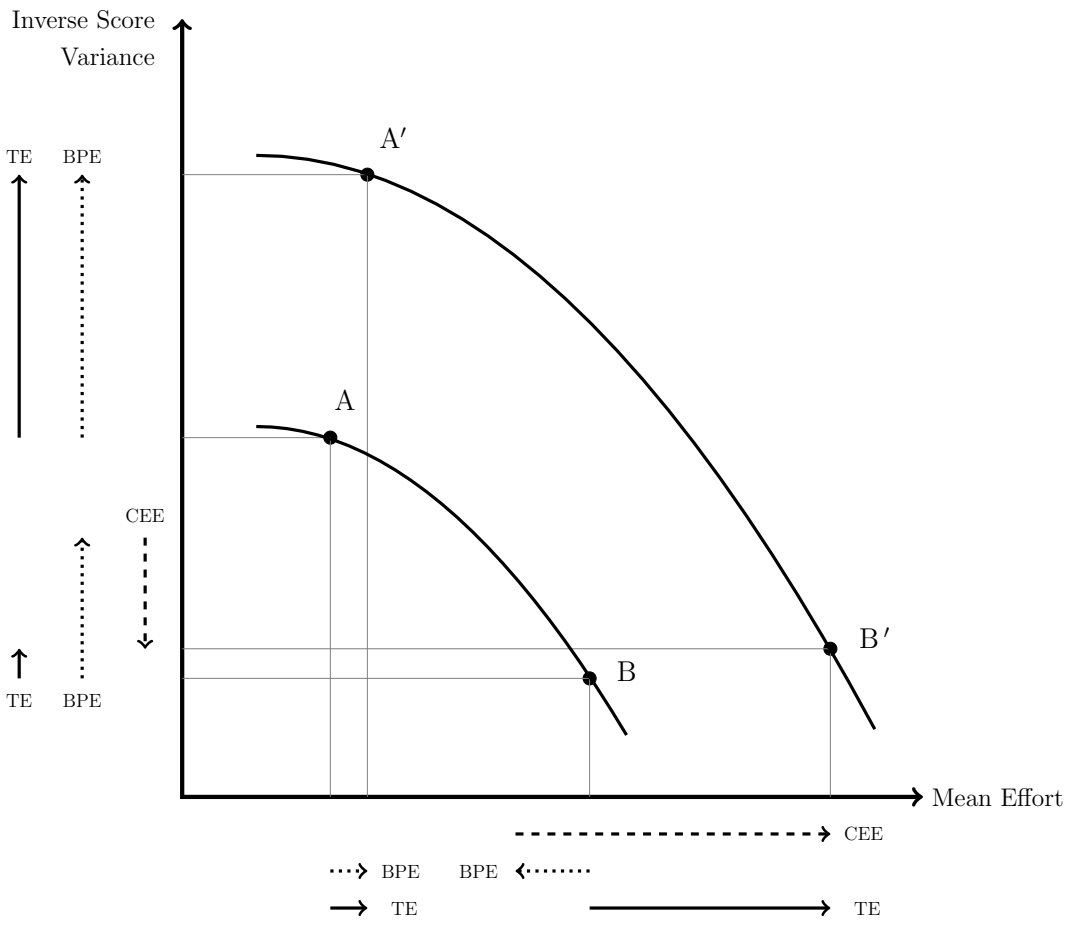
	(1)	(2)	(3)	(4)
Target	Mean Effort		Inverse of the Test Score Variance	
(Percentile Position)	Bonus Payment Effect	Cost-Equating Effect	Bonus Payment	Cost-Equating Effect
5	0.18	0.00	0.0037	0.0000
19	-0.05	0.29	0.0037	0.0011
39	-0.65	0.85	0.0031	0.0020
50	-0.99	1.23	0.0031	0.0018
76	-2.27	2.82	0.0047	-0.0016
95	-3.08	3.65	0.0071	-0.0061

In columns (1) and (2), we present the BPE and the CEE on mean effort, respectively, that occur in Figure 11 when we shift out from the homogeneous bonus payments frontier to the frontier associated with the bonus scheme that attaches more weight to low-performing students. In columns (3) and (4), we do the same but report the DE and CEE on the inverse of the test score variance.

transition from point B to point B').

In Figure 12, the transition from point A to point A' illustrates only the BPE, showing that, holding the target fixed at the real NCLB target and switching regimes from homogenous to heterogenous bonus payments increases mean effort. The precise increase is 0.18 developmental scale points, as shown in column (1) and Table 7. The increase results from the new bonus payment regime assigning the most weight to low-performing students, essentially 'doubling up' on already strong incentives for those students. For example, the mean effort gain (from switching bonus payment regimes) among students below the median of the predicted score distribution is 0.84 developmental scale points (12 percent of a standard deviation). In contrast, students above the median lose 0.47 developmental scale points (on average), implying that the decline in effort among those at the top of the distribution is not high enough to offset the gains at the bottom (because incentives for high-performing students students were quite low initially). In addition, because students at the bottom get disproportionately more effort, there is a decrease in the test score variance. Therefore, for relatively low proficiency targets, only the BPE is needed to generate the shift out of the frontier.[51]

[51]That is, to simultaneously cause an increase mean effort and reduction in test score variance.



Notes: In this figure, we demonstrate the forces that cause the frontier under the regime in which bonus payments are decreasing in the predicted score to shift out relative to the homogeneous b frontier in panel (a). We illustrate these forces at two specific points: a fixed target at the 5th percentile of the predicted score distribution (point A to point A') and a fixed target at the 95th percentile of the predicted score distribution (point B point B'). The label "BPE" describes the bonus payment effect on mean effort and inverse test score variance that arises from changing the the bonus payment structure to attach more weight (i.e, a higher bonus) to relatively low-performing students. The label "CEE" describes the cost-equating effect, which is the amount that mean effort and the inverse test score variance change as a result of further altering the bonus payment (by the same factor for all students) to achieve cost-equivalence across target regimes. The label "TE" describes the total effect, calculated as the sum of the bonus payment effect and the cost-equating effect.

FIGURE 12 – INTUITION FOR FIXED FRONTIERS WITH HETEROGENEOUS BONUS PAYMENTS

For all other targets reported on the frontier, the decompositions in Table 7 show that the CEE is needed to increase mean effort when switching bonus payment regimes, as the BPE alone results in a mean effort reduction. To illustrate why this is the case, refer to Figure 12 and suppose we instead start at the point B and consider the shift to B'. At relatively high proficiency targets such as this one,[52] the BPE creates a tension between the incentive to devote effort to high-performing students and the incentive to devote effort to low-performing students due to the heterogeneous bonus payments.

As a result of the BPE, high-performing students are thus allocated less effort than under the homogenous bonus payment regime, whereas low-performing students receive more, leading to a decrease in test score variance (an increase in inverse variance). For example, students with predicted scores above the median receive 6.96 developmental scale points less effort (on average) as a result of the BPE, while students with predicted scores below the median receive 0.93 developmental scale points more effort. These effects are equivalent to 0.97 and 0.13 standard deviations of the test score, respectively. The increase in effort among low-performing students is therefore not high enough to compensate for the loss among high-performing students, leading to an overall reduction in average effort.

Now consider the need to maintain cost equivalence and the CEE. Lower (unadjusted) mean effort under the new regime implies that costs are too low. Thus, in order to equate costs with the NCLB benchmark, a higher bonus payment must be offered to increase teacher effort and the proficiency rate. Doing so increases mean effort, and the test score variance also increases (inverse variance decreases), as high-performing students benefit disproportionately from the higher bonus payment, again owing to their marginal position in the incentive strength distribution.

[52] More specifically, the BPE results in a mean effort increases for targets up to the 11th percentile of the predicted score distribution, after which point the CEE is needed to increase mean effort and generate the frontier's outward shift. Prior to targets at the 11th percentile, the BPE alone can generate the shift of the frontier.

In this case, the CEE increases mean effort among students with predicted scores above the median by 6.57 developmental scale points (0.92 standard deviations) and by only 0.63 scale points (0.09 standard deviations) for students with predicted scores below the median.

Applying similar reasoning to other points on the homogenous bonus payment frontier results in the same outward shift, thus tracing out the full frontier under the heterogeneous bonus payment regime that pays a greater reward for low-performing students.

As Figure 11 makes very clear, the scheme that assigns more weight to high-performing students is dominated by both other regimes. For brevity, we do not provide a full decomposition of the BPE and CEE that occur when switching from the homogeneous bonus payments regime to the regime that attaches more weight to high performers. Instead, we provide an overview of the adjustment that occurs.

As above, the mechanics of the adjustment depend on where the proficiency target is located. When the proficiency target is relatively low, it presents teachers with strong incentives to devote effort to low-performing students but the heterogeneous bonus payments provide strong incentives to devote effort to high-performing students. Low-performing students are thus allocated less effort than under the homogeneous bonus payment regime, whereas high-performing students receive more, leading to a increase in test score variance (a reduction in inverse variance). For example, at the real NCLB target, students below the median of the predicted score distribution lose 1.07 developmental scale points worth of effort, while those above the median gain 0.55 developmental scale points, implying that the increase in effort among high-performing students is not high enough to compensate for the loss among low-performing students, leading to a reduction in overall effort.

At high proficiency targets, both proficiency target incentives and bonus payment incentives are strongest for students who are high in the predicted test score distribution. These students receive the largest amount of extra effort, while low-

performing students experience the largest reduction. When the target is set at the 95th percentile, for example, students in the top quartile of the predicted score distribution gain 1.40 developmental scale points and students in the bottom three quartiles lose 1.06 scale points, leading to an overall reduction in mean effort.[53] Since the strongest students experience test score gains and weakest experience losses, inequality (test score variance) also rises. Together, the effects on mean effort and test score variance result in a frontier that is interior to the frontiers of the other two regimes.

Test Score Gaps Across Demographic Groups

A key finding to emerge from our analysis is that the regime offering greater bonus payments for low-performing students dominates the homogenous bonus payment regime for reasons we have just explained, both in terms of mean effort and test score variance. Further, and it is worth reiterating, this scheme costs the same as the incentive scheme that policymakers actually implemented.

The potential gains from switching to such a regime are substantial. One graphic way of highlighting the gains is to document the implied effects of the regime on test score gaps across student subgroups. Table 8 below reports three test score gaps that are of great interest to policymakers: the white-black test score gap, the gap between students of college-educated and non-college educated parents, and the gap between the 90th and 10th percentile of the test score distribution. For each test score gap, columns (1) and (2) show the observed gap in the data and the gap predicted by our model, respectively.

As our model predicts the observed gaps quite well, in column (3) we show the percentage of the predicted gap that can be eliminated by switching to the

[53] We do not cut the distribution at the median here (as we do in the illustrative examples above) because, at such a high target, even the group of students above the median of the predicted score distribution loses effort, *on average*, when switching to the regime that places more weight on high-performing students. It is only the top quartile of the predicted score distribution that benefits from the regime change at such a high proficiency target.

regime where bonus payments are higher for low-performing students. Redistributing bonus payments across students this way reduces the black-white test score by 11 percent of its original value *without changing overall costs*. The gap between children of college-educated and less than college-educated parents falls by a similarly substantial margin – by ten percentage points.

Table 8 – Test Score Gaps and Heterogeneous Bonus Payments

	(1)	(2)	(3)
			Fixed Target ($y^T = 247$ (5th Percentile))
Test Score Gap	Observed in Data (SD Units)	Predicted by Model (SD Units)	% of predicted gap eliminated with b decreasing in $\hat{y}$
White versus Black	0.78	0.77	11%
College-Educated versus less than College-Educated Parents	0.74	0.74	10%
90th versus 10th Percentile	2.55	2.49	9%

In columns (1) and (2), test score gaps are reported in (student-level) standard deviation units. In column (3), the percentage of the predicted gap (column 2) obtained by the heterogeneous bonus payment regime is evaluated using a the real NCLB fixed target of 247 developmental scale points, which is the fifth percentile of the predicted score distribution.

X.C. Value-Added Targets

We now explore the properties of value-added (VA) targets.[54] In doing so, we exploit the linkage between fixed and value-added targets, explained above.

Specifically, defining Y^{fixed} as the set of fixed targets we considered above when constructing the frontiers, and Ω as the set of VA multiplicative coefficients (see the Simulation Appendix), then for each fixed target in the set Y^{fixed}, we derive the corresponding VA target intercept δ for all $\alpha \in \Omega$. Specifically, for each α under consideration, we choose the corresponding level of δ to ensure the mean

[54] When we set a given VA target, we assume that all of the rules under NCLB continue to operate – there are many, relating to demographic subgroups, confidence intervals, 'safe harbour' provisions, etc. – with the important exception that test score proficiency targets are now made student-specific.

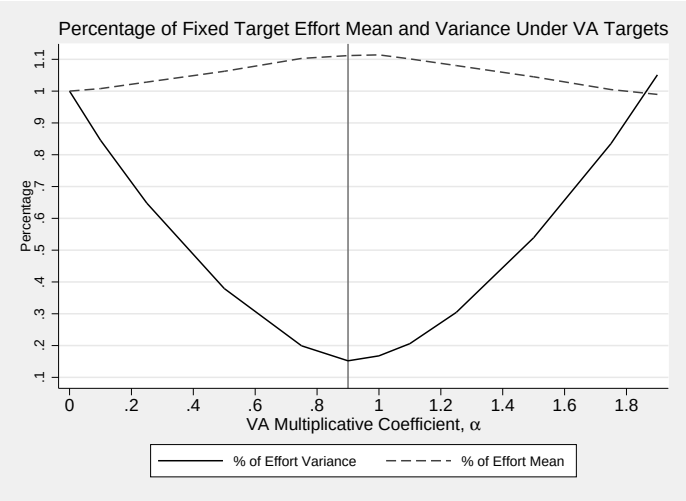
of the resulting incentive strength distribution is equivalent to the mean under the associated fixed target, the (α, δ) pairs giving the twelve different VA targets we examine for each fixed target in Y^{fixed}.

Our interest centers on how outcomes change as we vary the VA multiplicative coefficient, α, in order to incorporate more student-specific information into the target (through the use of the prior score). Two general properties emerges from our analysis of VA targets: over an extensive target range, VA targets result in less effort variance than their fixed target counterparts and in at least as much mean effort.

To illustrate the effort variance property, we note that, relative to fixed targets (in which $\alpha = 0$), increasing α up to a critical value α^* reduces the variance of incentive strength, causing teachers to exert similar levels of effort toward all students. (In contrast, increasing α past α^* increases the variance of incentive strength, eventually leading to greater dispersion in effort than under fixed targets.) That critical value α^* is the coefficient from the linear regression of $\hat{y}_{it}$ on y_{t-1}, equal to $\frac{cov(\hat{y}_{it}, y_{i,t-1})}{var(y_{i,t-1})}$ (see Simulation Appendix), which is estimated to be 0.937 in our data. Conditioning proficiency targets on students' prior scores judiciously therefore allows policymakers to use VA targets to make a higher fraction of students marginal than under fixed targets, resulting in lower inequality in teacher effort across students.

Figure 13 shows that, compared to the fixed target baseline, increasing α toward α^* (depicted by the vertical line) reduces the variance in effort across students progressively, while increasing α above α^* increases the variance progressively, eventually resulting in higher variance than under fixed target regimes. At the variance-minimizing choice of α, VA targets result in less than 20 percent of the effort variance observed under fixed targets.

Figure 13 also shows VA targets deliver at least as much average effort as fixed targets, with mean effort under VA targets peaking at 110 percent of the value under

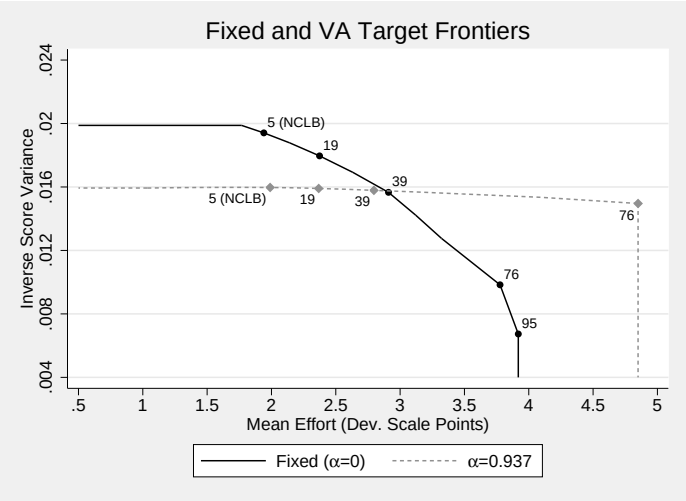


Notes: This figure shows the percentages of the fixed target effort mean and variance that are attained by VA targets with different multiplicative coefficients, α. The dashed line depicts the average fraction of the fixed target effort mean that is achieved by VA targets as a function of the VA target multiplicative coefficient, α. The solid line depicts the average fraction of the fixed target effort variance that is achieved by VA targets as a function of the VA target multiplicative coefficient, α. The vertical line depicts the effort-variance-minimizing α, equal to 0.937 in our data.

FIGURE 13 – PERCENTAGE OF FIXED TARGET MEAN AND VARIANCE ACHIEVED BY VA TARGETS

fixed targets when α is equal to α^*. To illustrate when VA targets dominate in terms of mean effort, consider Figure 14 below, which further demonstrates the effort mean and variance properties of VA targets in a frontier picture akin to Figure 9, while using the effort variance-minimizing value α^* as the multiplicative coefficient for the VA targets. The fixed target corresponding to each VA target is labelled on the VA frontier in Figure 14, again using the percentile position of the fixed target in the distribution of predicted scores as the label.

By setting $\alpha = \alpha^*$, the variance of incentive strength is minimized and is constant across all of the fixed targets. Because incentive strength is so similar across all students, there is very little resulting variance in effort across students as we change δ and shift the distribution of incentive strength. Figure 14 shows that a direct corollary is that the test score variance is also nearly constant across all fixed target counterparts, as the variance of effort does not change and all students get similar boosts to test scores under each fixed target mean.



Notes: The solid line depicts the frontier from Figure 9. Each point on the solid line reflects the mean effort and inverse test score variance that prevails under a given fixed target. These are calculated by using our model to determine effort decisions and the resulting test score distribution under each fixed target. The point labels correspond to the percentile position of the fixed target in the distribution of student predicted scores. The dashed line depicts the frontier that arises under the set VA targets with the multiplicative coefficient α equal to the effort-variance-minimizing value of 0.937. For each VA target, we choose the VA intercept δ such that the mean of the incentive strength distribution under the VA target matches the mean of the incentive strength distribution under a given fixed target. The point labels on the dashed line correspond to the percentile position of the fixed target whose incentive strength mean the VA intercept δ is chosen to match.

FIGURE 14 – FIXED AND VA TARGET FRONTIERS

Comparing the fixed and VA frontiers in Figure 14 shows that the test score variance is higher (the inverse variance is lower) under VA targets than under fixed targets when VA targets match incentive strength means of fixed targets that are lower than the thirty-ninth percentile of the predicted score distribution – for example, the VA scheme results in 22 percent greater test score variance at the real NCLB fixed target. At these relatively low fixed targets, the wider variance of incentive strength under fixed regimes provides stronger incentives (than VA regimes) to redistribute effort to the lower tail of the predicted score distribution, disproportionately boosting scores of low-performing students and reducing test score variance.

For higher fixed targets, the test score variance under VA targets is lower (inverse variance is higher) than under the fixed target counterparts because the fixed target regimes result in more effort being allocated to relatively high-performing students. At higher fixed targets, VA schemes also result in greater average effort – for example, at the fixed target positioned at the seventy-sixth percentile of the

predicted score distribution, the VA target results in 28 percent greater mean effort and 34 percent less test score variance. This follows from the tight incentive strength distribution under VA targets, which implies that a larger mass of students have a reasonable chance of achieving proficiency than under the fixed target, enticing teachers to exert more effort.[55]

VA targets therefore outperform fixed targets (in terms of mean effort and test score variance) when policymakers set a relatively high proficiency threshold. In these cases, using student prior scores to narrow the incentive strength distribution results in both greater average effort and less test score inequality.

XI. Conclusion

This paper has made three related contributions. First, it set out a transparent semi-parametric approach for identifying the impact of incentives on effort. We used exogenous incentive variation associated with the introduction of a prominent accountability reform to identify the effort response of North Carolina teachers based on changes in test scores. Our approach rests on minimal assumptions, is easy to implement, and can be applied in other contexts to identify effort (detailed data and appropriate policy variation permitting) – valuable given that effort is typically unobserved and thus difficult to pin down.

Second, we developed a structural procedure based on a model of teacher effort setting, allowing us to identify the primitives underlying the effort response. Estimates of the model show that within-classroom tradeoffs in effort across students are important.

[55]We cannot trace out the VA frontier further out to compare VA targets with fixed targets that are higher than the seventy-sixth percentile of the predicted score distribution. Because the incentive strength distribution is very narrow when $\alpha = 0.937$, after the seventy-sixth percentile, the entire incentive strength distribution is too far below zero (or the proficiency threshold), implying that virtually no student has a reasonable chance at achieving proficiency. Regimes with fixed target means greater than this threshold are therefore unreasonable incentive schemes to put in place and our model correspondingly produces unreasonable results.

Third, the model and estimates then formed the basis of a counterfactual framework for measuring the performance of different incentive schemes on a comparable basis for the first time. The framework allows us to assess how effort changes with counterfactual incentives, and to compute the full distribution of scores under counterfactual incentive provisions. This serves as a valuable design tool at a time when states are re-visiting incentives under NCLB.

We used the framework to compare the performance of alternative incentive schemes, including those yet to be implemented, having placed them all on a common footing by equating costs. Three main findings emerge, each relevant to incentive design in education. First, we show that fixed targets (of the form taken by NCLB) give rise to a quantitatively significant tradeoff between teacher effort and student test score inequality. Second, their performance can be improved markedly by introducing student-specific *bonuses* that attach higher weight to low-performing students. We show that these can reduce the black-white test score gap by 11 percent and the score gap between children of college educated versus non-college educated parents by 10 percent, in each case at no extra cost. Third, switching from fixed to student-specific *targets* allows policymakers to reduce inequality in teacher effort across students by as much as 80 percent without any sacrifice in aggregate effort.

In related work, we are examining how the exogenous incentive variation in this study can be used to shed light on the nature of the underlying production technology in education. Building on our strategy for recovering unobserved effort, we explore how various education inputs, including teacher effort, persist. Such persistence effects are potentially very relevant for policy, speaking (among other things) to the issue of 'teaching to the test.'

References

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2005. "Social Preferences and the Response to Incentives: Evidence from Personnel Data." *Quarterly Journal of Economics*, 120(3): 917-962.

Burgess, Simon, Carol Propper, Helen Slater, and Deborah Wilson. 2005. "Who Wins and Who Loses from School Accountability? The Distribution of Educational Gain in English Secondary Schools." CEPR Discussion Paper No. 5248, September.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review*, 104(9): 2593-2632.

Copeland, Adam and Cyril Monnet. 2009. "The Welfare Effects of Incentive Schemes." *Review of Economic Studies*, 76(1): 93-113.

Cullen, Julie and Randall Reback. 2006. "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System" in *Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics*, edited by T. Gronberg and D. Jansen, Volume 14, Amsterdam: Elsevier Science.

Dee, Thomas S. and Brian Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management.* 30(3): 418-446.

Deming, David J., Sarah Cohodes, Jennifer Jennings, Christopher Jencks, and Maya Lopuch. 2013. "School Accountability, Postsecondary Attainment and Earnings." National Bureau of Economic Research Working Paper 19444.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review*, 101(5): 1739-74.

Figlio, David N. and Lawrence W. Kenny. 2007. "Individual Teacher Incentives and Student Performance." *Journal of Public Economics*, 91(5-6): 901-914.

Hoxby, Caroline M. 2002. "The Cost of Accountability." National Bureau of Economic Research Working Paper 8855.

Imberman, Scott and Michael Lovenheim. 2015. "Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System." *Review of Economics and Statistics*, 97(2): 364-86.

Laffont, Jean-Jacques, and Jean Tirole. 1993. *A Theory of Incentives in Procurement and Regulation*, MIT Press, Cambridge, MA.

Lavy, Victor 2009. "Performance Pay and Teachers' Effort, Productivity, and Grading

Ethics." *American Economic Review*, 99(5): 1979-2011.

Lazear, Edward P. 2000. "Performance Pay and Productivity." *American Economic Review*, 90(5): 1346-1361.

Macartney, Hugh. 2016. "The Dynamic Effects of Educational Accountability." *Journal of Labor Economics*, 34(1): 1-28.

Mas, Alexandre, and Enrico Moretti. 2009. "Peers at Work." *American Economic Review*, 99(1): 112-45.

Mirrlees, James A. 1975. "The Theory of Moral Hazard and Unobservable Behaviour: Part I." Mimeo, Oxford University. Reprinted in 1999, *Review of Economic Studies*, 66: 3-21.

Misra, Sanjog and Harikesh S. Nair. 2011. "A Structural Model of Sales-force Compensation Dynamics: Estimation and Field Implementation." *Quantitative Marketing and Economics*, 9(3): 211-257.

Neal, Derek and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-based Accountability." *Review of Economics and Statistics*, 92(2): 263-283.

Reback, Randall. 2008. "Teaching to the Rating: School Accountability and the Distribution of Student Achievement." *Journal of Public Economics*, 92(5-6): 1394-1415.

Reback, Randall, Jonah Rockoff, and Heather L. Schwartz. 2011. "Under Pressure: Job Security, Resource Allocation, and Productivity in Schools Under NCLB." National Bureau of Economic Research Working Paper 16745.

Rivkin, Steven G., Eric A. Hanushek and John T. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica*, 73(2): 417-458.

Appendices

A. Illustrative Model Appendix

This appendix presents a stylized model of the education process that accords with the simple motivating example discussed in Section II.

The stylized model links accountability incentives to outcomes via discretionary action, 'effort,' taken to be changes in observable test scores that are attributable to incentive variation. This yields an effort function that depends on the parameters of the incentive scheme and, under threshold targets, a measure of incentive strength. The model serves as a means for analyzing the determinants of optimal effort.

For simplicity, we focus on the effort decision of a single teacher, teaching a single student (whose ability is allowed to change exogenously). Thus it describes an individual tutoring problem, unlike the structural model in the main analysis, which involves a given teacher choosing an optimal effort vector for an entire classroom of students.

The stylized model has three elements:

First, there is a *test score technology* relating measured education output y to various inputs. We write this as $y_i = q(e_i, \theta_i) + \epsilon_i$, where y_i is the score outcome for student i; $q(e_i, \theta_i)$ is a systematic component – student i's predicted score, depending on teacher effort (e_i) and student ability (θ_i); and ϵ_i is random additive noise that affects scores. We define $H(\cdot)$ and $h(\cdot)$ as the cumulative distribution and probability density functions of the negative of this noise, $-\epsilon_i$, and assume these functions are common across all students.

Second, we characterize an *incentive scheme* by a target y^T and a reward b, both exogenously given.

Third, the teacher faces a convex *cost of effort*. We write this as $c(e_i)$ for student i.

Taking these elements together, we can write down the teacher's objective when teaching student i under a threshold-based scheme as $U_i = b \cdot 1_{y_i \geq y_i^T} - c(e_i)$, which in expectation is given by $b \cdot \Pr[q(e_i, \theta_i) - y_i^T \geq -\epsilon_i] - c(e_i) = b \cdot H[q(e_i, \theta_i) - y_i^T] - c(e_i)$.

Optimal effort e_i^* will then implicitly satisfy the first-order condition, given by

$$b \cdot h[q(e_i, \theta_i) - y_i^T] \frac{\partial q(e_i, \theta_i)}{\partial e_i} = c'(e_i). \tag{13}$$

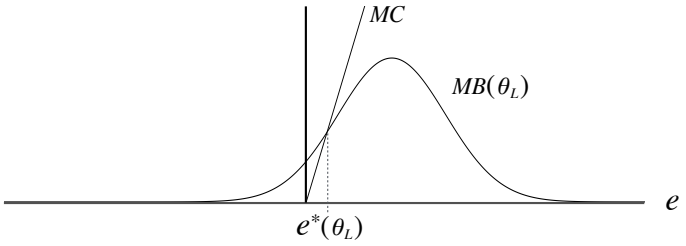
Unlike a piece rate, optimal effort is a function of the target. Further, it depends on the value of θ in a systematic way – a point we now develop.

Take the case where $q(e, \theta)$ is simply the sum of its arguments: $q(e_i, \theta_i) = e_i + \theta_i$. The additively separable assumption implies that the marginal benefit of effort, given by the LHS of (13), simplifies to $b \cdot h[\theta_i + e_i - y_i^T]$. Holding the reward parameter b fixed, marginal benefit is then a function of two quantities. The first is the gap between the systematic component of the score, $q(e_i, \theta_i)$ and the target for student i; the second is the density of the error in the performance measure, $h(\cdot)$, evaluated at that gap.

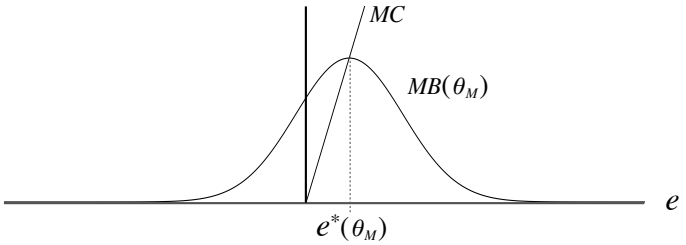
For illustration, suppose that the error term is unimodal, peaking at a mean of zero. Then consider three cases, corresponding to variation in the underlying conditions governing education production for three types of student who vary by ability (low-, moderate- and high-ability), where $\theta_L < \theta_M < \theta_H$, respectively.

In Figure A.1, we illustrate the effects of shifting θ on optimal effort, found at the intersection of the marginal cost and marginal benefit curves. Effort is on the horizontal axis, and the intersection with the vertical axis indicates zero effort – the origin for the marginal cost of effort curve in each panel. The peak of the marginal benefit curve will be found at the effort level, $\bar{e}$, for which the predicted score equals the target – we assume a symmetric distribution in the figure. Taking the target to be fixed at the same value across all three panels, in the linear case we have $\bar{e}(\theta) = y^T - \theta$, which is declining in the underlying conditions θ, leading the marginal benefit curve to shift left as underlying 'production' conditions become more favorable.

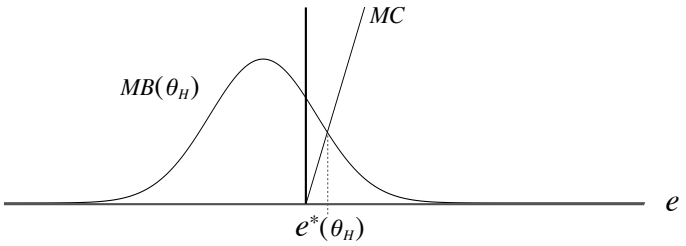
Consider the teacher's effort choice problem in the first case, where $\theta = \theta_L$. We illustrate this case in panel (a). The marginal benefit curve, conditioning on θ_L, will simply be the product of (fixed) b and the density $h(\cdot)$, tracing out the shape of the latter. Optimal effort, $e^*(\theta_L)$, is determined by the intersection of this marginal benefit curve and the given marginal cost curve. Taking the target as fixed, the low value of θ makes it



(a) Low θ relative to y^T



(b) Intermediate θ relative to y^T



(c) High θ relative to y^T

FIGURE A.1 – OPTIMAL EFFORT AND VARYING EXOGENOUS PRODUCTION CONDITIONS UNDER A THRESHOLD SCHEME

very unlikely that the student will exceed her performance target, even if effort is set at a high level; thus, the incentive to exert costly effort will be correspondingly low.

It is straightforward to see how optimal effort changes as we raise the θ parameter. Shifting from θ_L to θ_M, the marginal benefit curve moves to the left in panel (b), in turn moving the intersection between marginal benefit and marginal cost to the right (at least in this intermediate case). Intuitively, the underlying production conditions relative to the target make effort more productive in terms of raising the odds of exceeding the target, so the teacher will have an incentive to exert higher effort. This incentive is unlikely to be monotonic, however. Panel (c) illustrates the case where $\theta = \theta_H$, the underlying

production conditions being *so* favorable that the teacher is likely to satisfy the target even while exerting little effort. Where marginal cost and marginal benefit intersect, the height of the marginal benefit curve is relatively low, reflecting the low marginal productivity of effort. This in turn leads to a low level of effort, lower than the case where $\theta = \theta_M$.

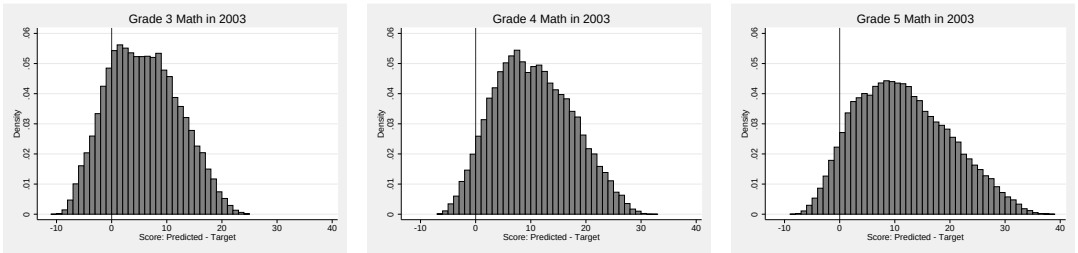
Generalizing, the optimal effort function will depend on incentive strength (the gap between the target and the student's predicted score). Further, under a threshold scheme, the function should follow an inverted-U, peaking where incentives are strongest.

B. Robustness Check and Ruling out A Rival Story

B.1. Validity of the Reduced-Form Approach: Testing for Bunching

When presenting the research design, we drew attention to the required exogeneity of the incentive 'shock.' Indirect light can be shed on this by examining bunching in the distributions of the predicted *ex ante* incentive strength measures, especially in the vicinity of the target.

To give a sense of the grade-specific distributions of our ex ante incentive strength measure that emerge from applying the proposed recipe, Figure B.1 plots the incentive-strength distributions for Grades 3, 4, and 5 mathematics in 2003. We are especially interested to see if NCLB produces any bunching around the relevant target.



(a) Grade 3 (b) Grade 4 (c) Grade 5

Figure B.1 – Distribution of Predicted Scores minus the NCLB Target

In each of the panels, the fixed NCLB target occurs at zero, as indicated by the vertical line. Based on the distribution of predicted scores, the figure provides no evidence of

bunching. This lends support to the notion that the NCLB 'shock' was indeed exogenous, affecting the effort of educators but not other determinants of student test scores.

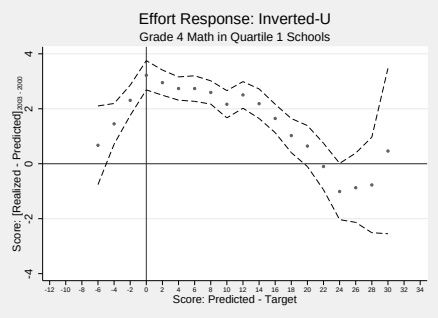
B.2. Response to the Target and Not Position in School-Specific Distribution

Our maintained hypothesis is that we are uncovering an effort response with respect to the incentive strength measure, π. As an alternative, effort might vary with respect to a student's relative position in the predicted score ($\hat{y}$) distribution within his or her school. For example, it is possible that educators responded to NCLB by targeting effort towards students at a particular point of the $\hat{y}$ distribution and that this point happened to coincide with the value of $\hat{y}$ where π under NCLB was close to zero. Such a response is in the spirit of Duflo, Dupas, and Kremer (2011), who set out a model in which teachers choose a particular type (or quality) of effort such that students at a certain point in the ability distribution will benefit most. Students who are further away from this point require a different type of effort or teaching style, so they do not benefit as much and may even perform worse than they otherwise would. If teachers in North Carolina responded to NCLB's introduction by tailoring teaching methods best-suited for students at the point in the ability distribution where π equalled zero, then varying π counterfactually to make inferences about competing accountability schemes would seem unwarranted.

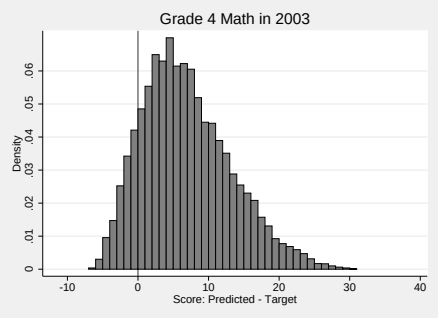
To assess this possibility, we exploit the richness of the administrative data – specifically, by determining the effort responses and corresponding π densities separately for four types of school. We divide schools according to the mean of their ex-ante predicted pass rates, and further, on the basis of which quartile (in terms of the predicted pass rate) they are in.[56] If schools responded to NCLB by tailoring effort toward a particular part of the ability distribution, we should observe the peak of the effort response shifting to the right as that point in the ability distribution shifts right across the types of school. This is not the case: Figure B.2 plots the effort responses and π densities separately for schools in each of the quartiles of the school-level (ex-ante) predicted pass rate. As one moves up the quartiles, the π distribution shifts rightward, implying that a student with

[56] A student is predicted to pass when $\pi = \hat{y} - y^T > 0$.

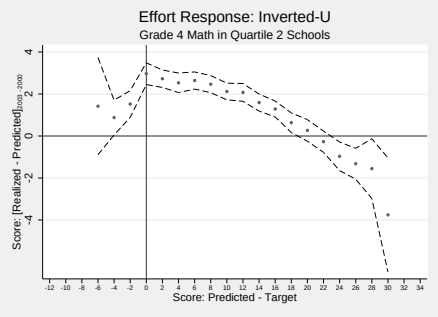
a value of π near zero in quartile-one schools will have a different relative position in the $\hat{y}$ distribution than a student with a value of π near zero in the quartile two, three or four schools. Yet the peak effort response occurs close to $\pi = 0$ and the effort function maintains a similar shape across each of the quartiles. This supports the view that schools respond to a student's proximity to the proficiency threshold and not his or her relative position in the predicted score distribution.



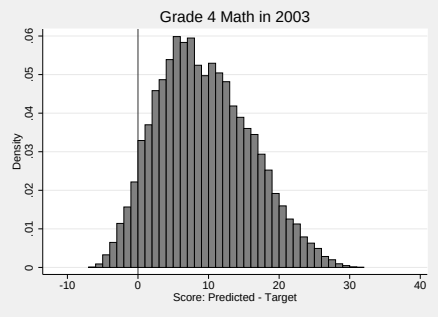
(a) Effort in Q1 Pass Rate Schools



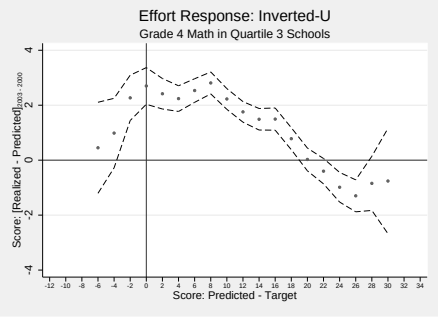
(b) π Density in Q1 Pass Rate Schools



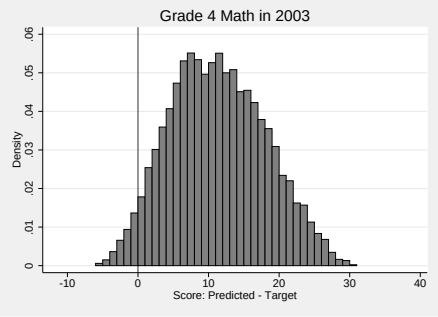
(c) Effort in Q2 Pass Rate Schools



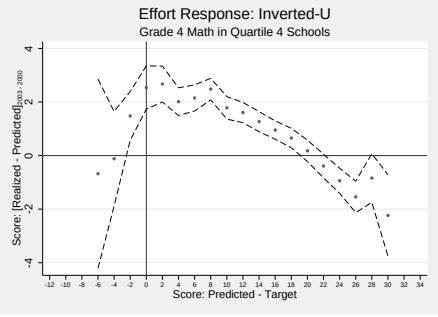
(d) π Density in Q2 Pass Rate Schools



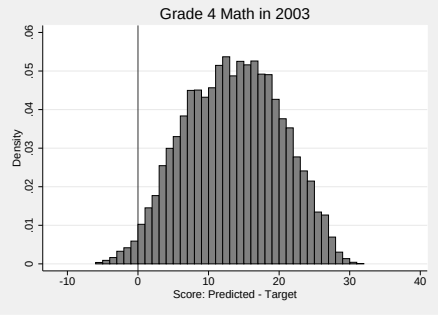
(e) Effort in Q3 Pass Rate Schools



(f) π Density in Q3 Pass Rate Schools



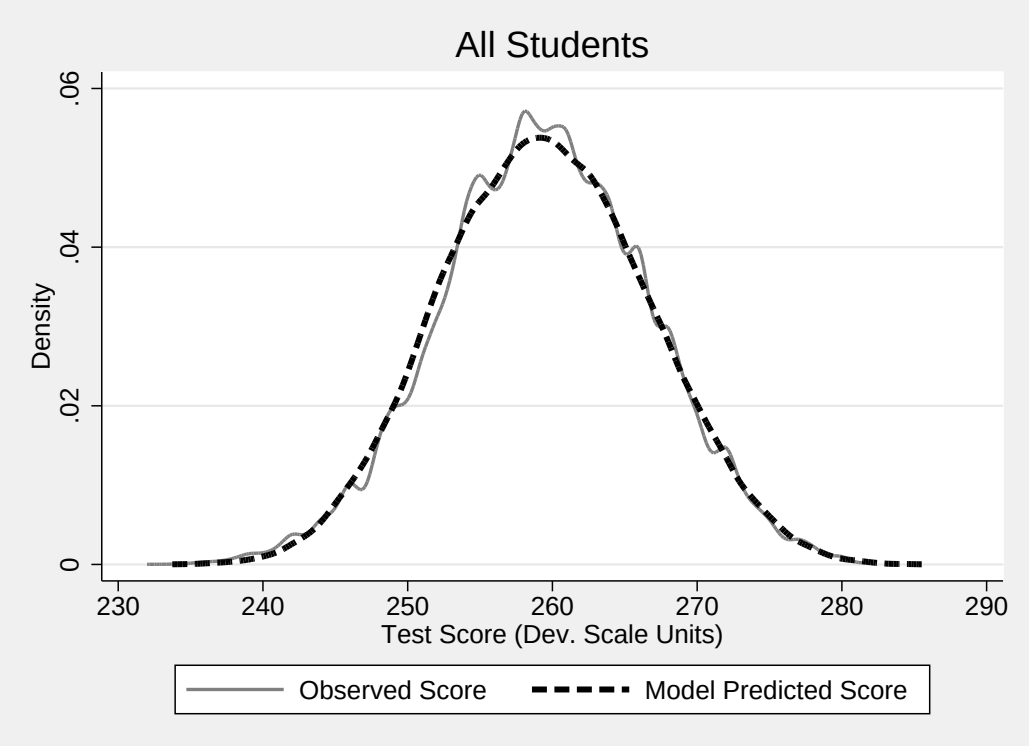
(g) Effort in Q4 Pass Rate Schools



(h) π Density in Q4 Pass Rate Schools

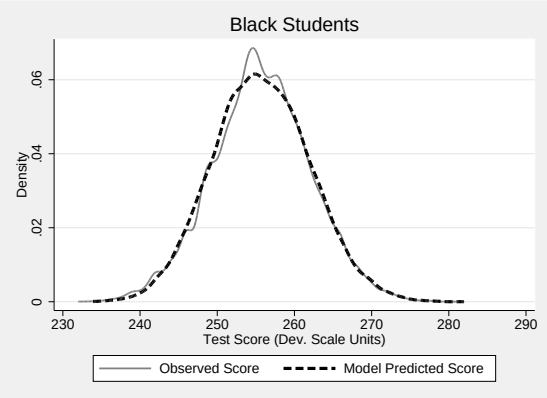
FIGURE B.2 – RESPONDING TO π RATHER THAN THE RELATIVE POSITION OF $\hat{y}$

C. Model Fit

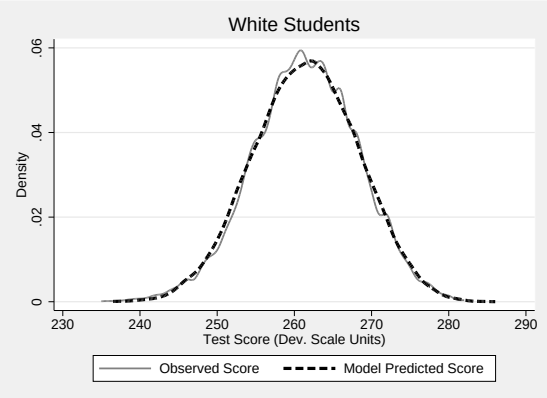


Notes: This figure presents the density of observed test scores (measured in developmental scale units) and the density of model-predicted test scores for the full sample.

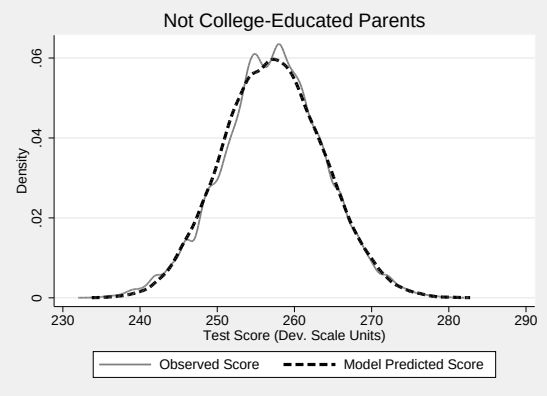
Figure C.1 – Distributions of Observed and Model Predicted Scores



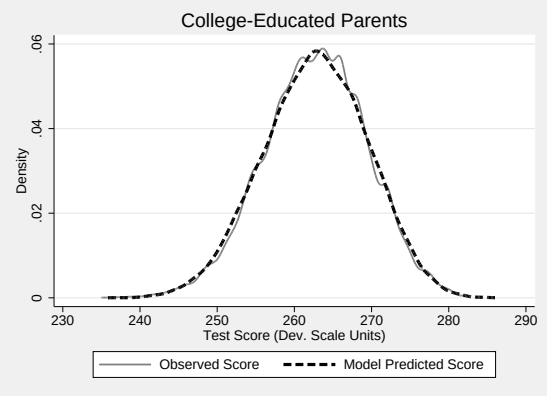
(a) Black Students



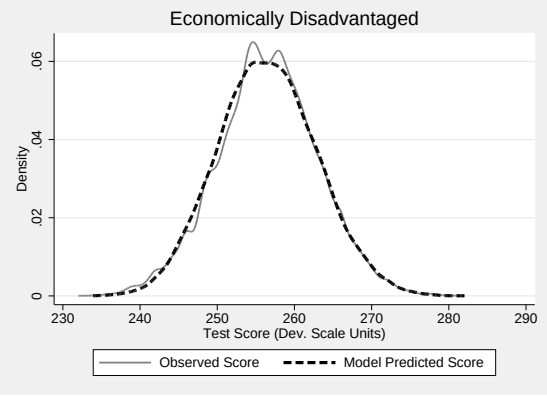
(b) White Students



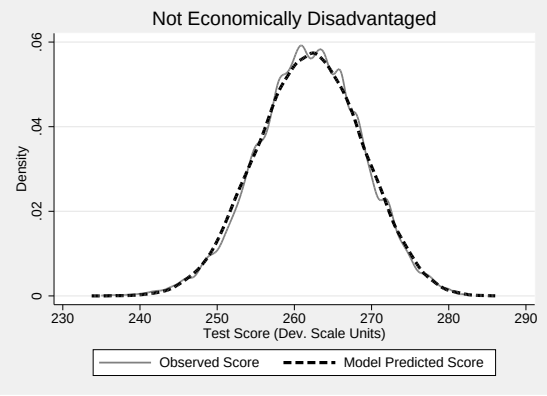
(c) Not College-Educated Parents



(d) College-Educated Parents



(e) Economically Disadvantaged



(f) Not Economically Disadvantaged

Notes: These figures present the density of observed test scores (measured in developmental scale units) and the density of model-predicted test scores for various sub-samples of students.

FIGURE C.2 – SUBGROUP DISTRIBUTIONS OF OBSERVED AND MODEL PREDICTED SCORES

D. Simulation Appendix

This appendix provides further details about target setting and the cost-equating procedures we use when carrying out the counterfactual simulations.

D.2. Counterfactual Fixed Targets

We construct counterfactual fixed targets measured on the End-of-Grade Mathematics test developmental scale, following the rules of NCLB. In particular, we consider fixed targets from 237 developmental scale points to 269 developmental scale points, increasing the target by an increment of 2 points on each iteration; for convenience, the set of fixed targets is defined as

$$Y^{fixed} = \{237, 239, 241, \ldots 247, \ldots 263, 265, 269\}. \tag{14}$$

This set of fixed targets covers the entirety of the predicted score distribution (aside from the very top). For completeness, Table D.1 shows the mapping between each of the developmental scale point targets we consider and the corresponding percentiles in the predicted score distribution (of $\hat{y}$). The true NCLB test score proficiency target (in bold in the table) is set at 247 developmental scale points, which corresponds to the fifth percentile of the predicted score distribution.

D.3. Counterfactual VA Targets

To keep the analysis tractable, we restrict attention to VA targets that use students' prior scores from only one subject (mathematics) and are linear in those scores.[57] Specifically, we allow the prior score to play a progressively more important role by varying α in the set $\Omega = \{0, 0.1, 0.25, 0.5, 0.75, 0.9, 1, 1.1, 1.25, 1.5, 1.75, 1.9\}$. For each $\alpha \in \Omega$, we then

[57] To qualify as a VA, a target must simply use information contained in students' prior scores. As such, there are many potential sets of VA targets. For example, during the 1990s, North Carolina's own ABCs program used *both* prior math and reading scores in in a linear way when setting targets for either subject while South Carolina's accountability program used both scores and incorporated linear, quadratic, and interacted terms.

Table D.1 – Developmental Scale Point Targets and Corresponding Percentiles

Developmental Scale Point Target	Percentile in Predicted Score Distribution
237	-
239	1
241	1
243	1
245	2
247	**5**
249	11
251	19
253	29
255	39
257	49
259	59
261	68
263	76
265	84
267	90
269	95

select the intercept of the VA target δ, so that the mean of student VA targets matches the value (in developmental scale points) of a given fixed target in the set Y^{fixed}. For example, suppose we are matching real NCLB fixed target, 247. In that case, we have $\delta = 247 - \alpha\bar{y}_{t-1}$, which implies that the mean of the VA targets is also 247. For a given $\alpha \in \Omega$, we conduct this exercise for each fixed target in Y^{fixed}.

Matching the mean of VA targets to fixed targets ensures that each fixed target has a VA counterpart that delivers the same mean for the student incentive strength distribution, $\hat{y} - y_{it}^T$, as the fixed target does. The VA counterparts, however, will generally have smaller variances because the use of the prior score allows us to set more appropriate student-specific targets, thus making many more students marginal. By considering several different multiplicative coefficients, α, and adjusting the intercept, δ, to match the means of the full range of fixed targets, we are able to explore the effects on student outcomes of both mean shifts of, and variance changes to, the incentive strength distribution.

Incentive Strength Variance-Minimizing α

Let $var(\hat{y}_{it} - y_{it}^T)$ denote the variance of incentive strength across all students for any

given target y_{it}^T. For VA targets, we have $y_{it}^T = \delta + \alpha y_{it-1},\ \forall i$, which allows us to write

$$var(\hat{y}_{it} - y_{it}^T) = var(\hat{y}_{it}) \ + \ \alpha[\alpha var(y_{it-1}) - 2cov(\hat{y}_{it}, y_{it-1})],$$

where $cov(x, z)$ denotes the covariance between two random variables x and z.[58] Taking the partial derivative with respect to α and setting it equal to zero shows that the variance is minimized when

$$\alpha = \frac{cov(\hat{y}_{it}, y_{it-1})}{var(y_{it-1})} \equiv \alpha^*.$$

The variance is decreasing in α when $\alpha < \alpha^*$ and increasing in α when $\alpha > \alpha^*$.

D.4. Cost-Equating Procedure Under Homogenous Bonus Payments

Since the state must pay b for each student who is proficient, we can write the average cost under a set of counterfactual targets R as

$$Q_R = \frac{b \sum_{i=1}^{N_t} 1\left(\hat{y}_i + e^*(\hat{y}_i, y_{it}^R, \widehat{\mathbf{\Gamma}}; c) + \epsilon_i - y_i^R \geq 0\right)}{N_t}, \tag{15}$$

where N_t is the total number of students in the state, $\widehat{\mathbf{\Gamma}} \equiv [\widehat{(\frac{m}{b})} \cdot \frac{1}{w_i}, \widehat{\tilde{\mu}}, \widehat{\tilde{\sigma}}^2, \widehat{\theta}]$, and $\epsilon_i \sim N(0, \widehat{\sigma}^2)$. Note that the set R can be either from the family of fixed targets, in which case each student has the same target, $y_i^R = y^R,\ \forall\ i$, and y^R is an element of the set Y^{fixed} above, or R can be drawn from the family of VA targets, in which case each student has a student-specific target given by $y_{it}^R = \delta^R + \alpha^R y_{it-1}$.

To explain the cost-equating procedure clearly, we first focus on the case where bonus payments are constant across students and $w_i = 1$ for all students. While the parameter b is not separately identified from m in our model, we can (without loss of generality) normalize b to one and interpret the estimated ratio $\widehat{(\frac{m}{b})}$ accordingly.

To equate costs across target regimes, we define $b(k) = kb = k1 = k$ to be the original

[58] The parameter δ drops out because it is a non-varying constant that affects only the mean of incentive strength, not the variance.

bonus payment value multiplied by a constant $k > 0$. Computationally, multiplying b by k implies evaluating the effort function in equation (15) at the argument $\widehat{(\frac{m}{b})} \cdot \frac{1}{k}$ instead of $\widehat{(\frac{m}{b})}$ and multiplying the sum in the numerator by k (instead of b, which is normalized to one). We let Q^* denote the common average cost that all regimes must share, setting Q^* equal to the cost that prevails when our model is used to predict outcomes under the real NCLB fixed target of $y^T = 247$.[59]

With this notation in place, we use the following procedure to equate the cost that prevails under the target set R to Q^*: We first calculate the difference between the realized cost and the target cost, $Q_R - Q^*$. If the two costs are equivalent and the difference is zero, we stop. If they are different in absolute value, we adjust $b(k)$ by updating the value of k until $Q_R = Q^*$.

Changing k has two effects on average costs. The first effect is to change the amount paid per student who passes directly. This is seen by recognizing that the sum of the indicator variables in equation (15) is multiplied by a different value each time k adjusts. The second effect comes from the effect of changing k (equivalently, the bonus payment) on teacher effort decisions, which is made clear by the effort function in equation (15) being evaluated at the argument $\widehat{(\frac{m}{b})} \cdot \frac{1}{k}$ instead of $\widehat{(\frac{m}{b})}$. Increasing k therefore increases costs by raising both the payment per each passing student and enticing teachers to exert more effort, itself leading to more students reaching proficiency status. In contrast, decreasing k decreases costs by paying less per passing student and causing fewer students to pass (because teachers exert less effort).

D.5. Heterogeneous Bonus Payments

We consider two different regimes in which bonus payments are heterogenous across students.

In the first case, we model the student-specific bonus payment as $b^L(\hat{y}_i) = b\frac{(\hat{y}_{\max}+1-\hat{y}_i)}{\hat{y}_{\max}-\hat{y}_{\text{med}}+1}$, where $\hat{y}_{\max}$ is the maximum value of $\hat{y}_i$ across all students in the state and and $\hat{y}_{\text{med}}$

[59] In that case, the pass rate (average cost) is 0.9608, implying that just over 96 percent of fourth grade students were deemed proficient across the state. For comparison, the real pass rate in fourth grade in 2003 was also 0.96, implying that our model fits the data well and that this choice of Q^* reflects a cost policymakers are willing to pay.

is the median value of $\hat{y}_i$, implying that the bonus payment is greatest for the lowest-performing students (those with the lowest predicted scores). The parameter b is the per-student bonus payment from before, which is now scaled by the student-specific weight $w^L(\hat{y}_i) = \frac{(\hat{y}_{\max}+1-\hat{y}_i)}{\hat{y}_{\max}-\hat{y}_{\text{med}}+1}$. In second case, we model the student-specific bonus payment as $b^H(\hat{y}_i) = b\frac{(\hat{y}_i-y_{\min}+1)}{\hat{y}_{\text{med}}-\hat{y}_{\min}+1}$, where $\hat{y}_{\min}$ is the minimum value of $\hat{y}_i$ across all students in the state, implying that the bonus payment is greatest for the highest-performing students (those with the highest predicted scores $\hat{y}$). Here, the bonus payment b is scaled by the student-specific weight $w^H(\hat{y}_i) = \frac{(\hat{y}_i-y_{\min}+1)}{\hat{y}_{\text{med}}-\hat{y}_{\min}+1}$. To illustrate the form of these two heterogeneous bonus payment parameterizations, Figure D.1 below depicts each of them as a function of student predicted scores along with the baseline homogeneous bonus payment case in which $w_i = 1,\ \forall\ i$, and the density of the predicted score distribution.

These parameterizations of the heterogeneous bonus payment are convenient for three reasons. First, they illustrate two extremes: In the first case, the bonus is highest for the worst students and lowest for the best students; in the second case, the opposite is true. Second, they ensure that the original payment b is being multiplied by a number that ensures costs do not blow up: in the first case, b is multiplied by a value greater than 1 for students below the median and by a value less than 1 for students above the median: the reverse is true in the second case. (In both cases, the median student has b multiplied by 1.) Third, since the parameterizations are determined in a data-driven way, they can be calculated in any data set.

Heterogeneous Bonus Payments: Modifying the Cost-Equating Procedure

Under these heterogeneous bonus payment regimes, average costs are determined by

$$Q_R = \frac{b\sum_{i=1}^{N_t} w^L(\hat{y}_i)1\left(\hat{y}_i + e^*(\hat{y}_i, y_{it}^R, \mathbf{\Gamma}'; c) + \epsilon_i - y_i^R \geq 0\right)}{N_t} \tag{16}$$

and

$$Q_R = \frac{b \sum_{i=1}^{N_t} w^H(\hat{y}_i) 1\left(\hat{y}_i + e^*(\hat{y}_i, y_{it}^R, \mathbf{\Gamma}''; c) + \epsilon_i - y_i^R \geq 0\right)}{N_t} \tag{17}$$

where $\mathbf{\Gamma}' \equiv [(\frac{\widehat{m}}{b}) \cdot \frac{1}{w^L(\hat{y}_i)}, \widehat{\tilde{\mu}}, \widehat{\tilde{\sigma}}^2, \widehat{\theta}]$ and $\mathbf{\Gamma}'' \equiv [(\frac{\widehat{m}}{b}) \cdot \frac{1}{w^H(\hat{y}_i)}, \widehat{\tilde{\mu}}, \widehat{\tilde{\sigma}}^2, \widehat{\theta}]$ in the first and second cases, respectively. For each heterogeneous bonus payment case, we cost-equate across regimes to Q^* using the same methodology described above: we normalize b to 1, multiply it by k, and adjust k until costs equate to Q^*.

D.6. Non-Linear Production Technologies

Throughout our counterfactual simulations, we assume a linear production technology of the form

$$y_{it} = \hat{y}_{it} + e^*(\hat{y}_{it}, y_{it}^T, \widehat{\mathbf{\Gamma}}; c) + \epsilon_i, \tag{18}$$

in which student ability ($\hat{y}_{it}$) and teacher effort are additively separable. This technology gives rise to the following first-order condition for teacher effort setting:

$$b\tilde{f}\big(y^T - \hat{y}_{jt} - e_{jt}^*\big) = m\left[e_{jt}^* + \theta \sum_{i=1}^{N_c} e_i^*\right], \ \forall \ j = 1, \ldots, N_c. \tag{19}$$

If, instead, we allow for non-linear interactions between student ability and teacher effort, the production technology maybe be written as

$$y_{it} = \hat{y}_{it} + e^*(\hat{y}_{it}, y_{it}^T, \widehat{\mathbf{\Gamma}}; c) \ + \ \lambda \hat{y}_{it} \cdot e^*(\hat{y}_{it}, y_{it}^T, \widehat{\mathbf{\Gamma}}; c) + \epsilon_i, \tag{20}$$

where λ is the parameter governing the interaction of the two inputs. Using equation (20) as the production technology now implies the following first-order condition for teacher effort setting:

$$(21) \qquad b(1+\lambda\cdot\hat{y}_{jt})\tilde{f}\big(y^T-\hat{y}_{jt}-e^*_{jt}\big)=m\bigg[e^*_{jt}+\theta\sum_{i=1}^{N_c}e^*_i\bigg],\ \forall\ j=1,\ldots,N_c.$$

Notice that the first-order condition with a non-linear technology closely resembles the (implicit) first-order conditions under our heterogeneous bonus payment regimes (and assumed linear technology). In particular, when we assign more weight (higher bonus payments) to low-performing students, the first-order condition for teacher effort with a linear production technology is

$$(22) \qquad b\bigg(\frac{(\hat{y}_{\max}+1-\hat{y}_i)}{\hat{y}_{\max}-\hat{y}_{\text{med}}+1}\bigg)\tilde{f}\big(y^T-\hat{y}_{jt}-e^*_{jt}\big)=m\bigg[e^*_{jt}+\theta\sum_{i=1}^{N_c}e^*_i\bigg],\ \forall\ j=1,\ldots,N_c.$$

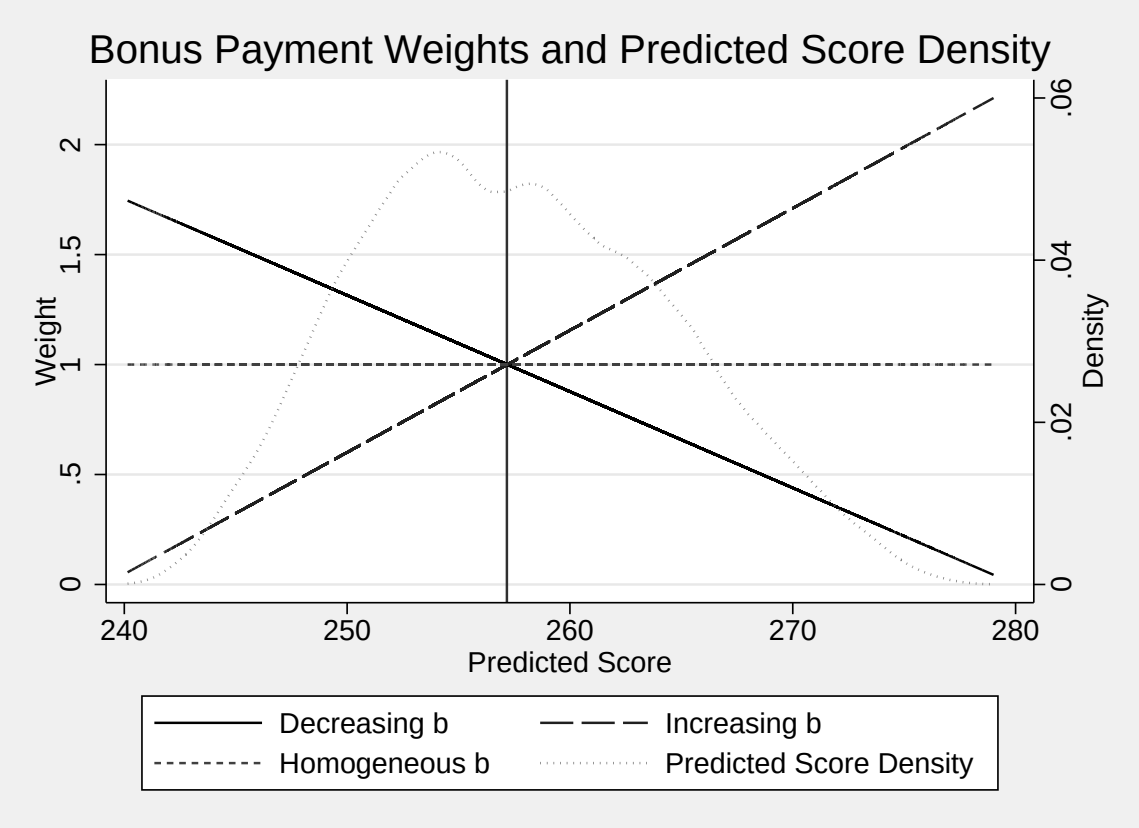
This heterogenous payment regime with a linear production technology is therefore akin to a special case of the non-linear production technology (with homogenous bonus payments) in which the interaction parameter $\lambda=-\frac{1}{\hat{y}_{\max}-\hat{y}_{\text{med}}+1}$. Likewise, the heterogenous bonus payment regime in which we assign more weight to high-performing students is akin to a special case of the non-linear production technology (with homogenous bonus payments) in which the interaction parameter $\lambda=\frac{1}{\hat{y}_{\text{med}}-\hat{y}_{\min}+1}$.

Therefore, although we do not explicitly estimate λ and run counterfactual simulations with a non-linear technology, our results from the heterogenous bonus payment cases are informative about the qualitative implications of such non-linear technologies. In particular, comparing the baseline simulation results (homogenous bonus payments with a linear production technology) to the results under the regime in which bonus payments are decreasing in student ability reflects how outcomes would change if the technology was non-linear and the interaction parameter λ was negative. In this case, relatively stronger incentives attach to low-performing students and the effort-inverse-variance frontier shifts outward, resulting in greater mean effort and less test score variance.

In contrast, comparing the baseline simulation results to the results under the regime in which bonus payments are increasing in student ability reflects how outcomes would change if the technology was non-linear and the interaction parameter λ was positive. In

this case, relatively stronger incentives attach to high-performing students and the effort-inverse-variance frontier shifts inward, resulting in less mean effort and more test score variance but still preserving the inherent tradeoff in proficiency target setting.

D.7. Figures



Notes: This figure shows the three bonus payment regimes as functions of student predicted scores and the density of the predicted score across all students. The dashed horizontal line shows the constant bonus payment case in which the bonus payment is normalized to one for all students. The decreasing solid line depicts the heterogeneous bonus payment regime in which we attach more weight to low-performing students. The increasing dashed line depicts the heterogeneous bonus payment regime in which we attach more weight to high-performing students. The dotted density profile shows the empirical density of the predicted score distribution across all students with the vertical line indicating the median value of the predicted score.

FIGURE D.1 – BONUS PAYMENT WEIGHTS AND PREDICTED SCORE DENSITY

E. Contribution of Research Design to Prior Literature

Given the incentive strength measure is very much related to, and builds upon, measures appearing in related prior work, it is worth drawing attention to seemingly subtle differences that turn out to be important in the development of our approach. First, our predicted student scores are based on *pre-reform* data – important in that we use these data to control for baseline effort (described shortly). Similar to the prediction algorithm in Reback (2008) and Deming, Cohodes, Jennings, Jencks and Lopuch (2013), we employ a flexible specification involving lagged test scores and several other student characteristics to calculate expected outcomes, though neither prior study has a pre-reform period.[60] Second, ours is a *continuous* measure, which we can compute for each student. In contrast, Deming *et al.* (2013) aggregate incentive strength to the school-level, and Neal and Schanzenbach (2010) group students into deciles of the ability distribution. The continuous measure is important when conducting counterfactuals, allowing us to evaluate how various targets change incentives throughout the student distribution.

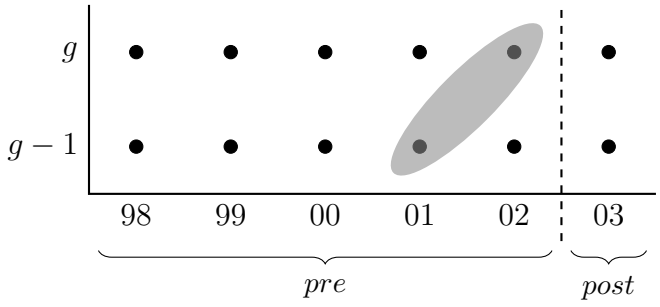
[60] Reback (2008) and Deming *et al.* (2013) both analyze the Texas accountability program that operated throughout the 1990's. Reback (2008) calculates a student-level passing probability, rather than a predicted score, as a measure of incentive strength.

F. Supplemental Tables and Figures

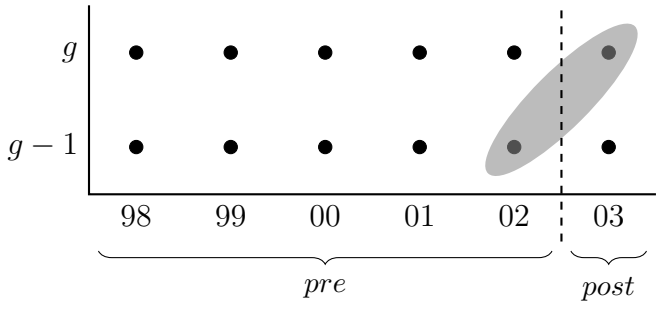
Table F.1 – Descriptive Statistics

	Student-Level		
	Mean	*Std. Dev.*	*N*
Performance Measures			
Math Score			
Grade 3	144.67	10.67	905,907
Grade 4	153.66	9.78	891,969
Grade 5	159.84	9.38	888,467
Grade 6	166.43	11.12	892,087
Grade 7	171.61	10.87	884,286
Grade 8	174.76	11.63	860,623
Math Growth			
Grade 3	13.85	6.30	841,720
Grade 4	9.40	5.96	730,627
Grade 5	6.82	5.29	733,037
Grade 6	7.55	5.68	722,491
Grade 7	5.99	5.60	718,994
Grade 8	3.73	5.86	705,095
Reading Score			
Grade 3	147.03	9.33	901,233
Grade 4	150.65	9.18	887,147
Grade 5	155.79	8.11	883,685
Grade 6	156.79	8.85	889,445
Grade 7	160.30	8.19	882,288
Grade 8	162.79	7.89	859,089
Reading Growth			
Grade 3	8.15	6.72	837,361
Grade 4	3.75	5.55	725,590
Grade 5	5.61	5.21	727,864
Grade 6	1.54	4.95	718,291
Grade 7	3.77	4.92	716,496
Grade 8	2.76	4.62	703,236
Demographics			
College-Educated Parents	0.27	0.44	5,456,948
Male	0.51	0.50	5,505,796
Minority	0.39	0.49	5,502,665
Disabled	0.14	0.35	5,498,312
Limited English Proficient	0.03	0.16	5,505,479
Free or Reduced-Price Lunch	0.42	0.49	3,947,605
	School-Level		
	Mean	*Std. Dev.*	*N*
Failed ABCs	0.27	0.45	14,052
Failed NCLB	0.37	0.48	5,014
Proficiency Rate	0.79	0.11	14,042

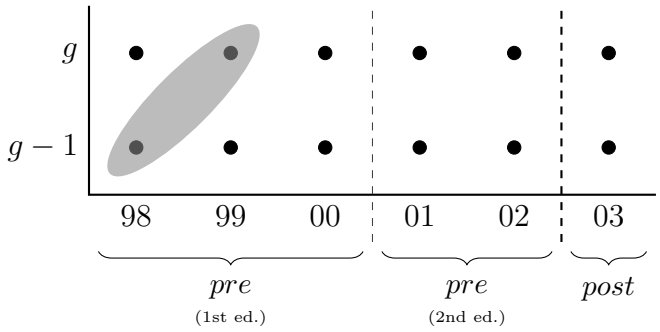
Notes: The sample excludes vocational, special education and alternative schools. We also exclude high schools and schools with a highest grade served lower than fifth grade. Student-level summary statistics are calculated over all third to eighth grade student-year observations from 1997-2005 in eligible schools. The free or reduced price lunch eligibility variable is not available prior to 1999. School-level summary statistics are calculated over all eligible school-year observations from 1997-2005. The NCLB performance indicator variable is not available prior to 2003, the year the program was introduced.



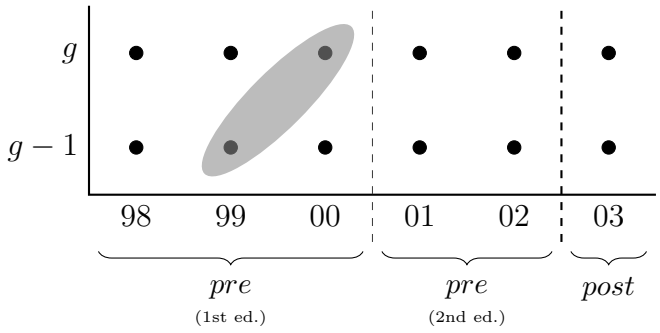
(a) First Step



(b) Second Step



(c) Third Step



(d) Fourth Step

Figure F.1 – Research Design in Pictures