Incentive Design in Education:
A Distributional Analysis[*]

by Hugh Macartney, Robert McMillan, and Uros Petronijevic[†]

October 2020

Abstract

This paper provides the first empirical analysis to examine how different incentive schemes in education affect the full distribution of student outcomes. Our starting point is an asymmetric hump-shaped relationship between teacher effort and incentive strength, estimated semi-parametrically using exogenous incentive variation and rich administrative data. We recover the primitives underlying this effort function, showing they can be identified by estimating a flexible effort-choice model. The model and resulting estimates are key inputs to a counterfactual framework we propose for tracing the unstudied effects of alternative accountability systems on the entire test score distribution, allowing effort to adjust endogenously. We find that widespread fixed-target systems give rise to a steep performance-inequality tradeoff; further, existing schemes can be modified to reduce test score inequality while improving average student performance at no extra cost. Overall, the combined estimation-simulation approach opens up new possibilities for incentive design in practice, as our new findings indicate.

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

JEL Classifications: D82, I21, J33, M52

[*] This is an updated version of NBER Working Paper 21835. We would like to thank Joseph Altonji, Peter Arcidiacono, David Deming, Giacomo De Giorgi, David Figlio, Chris Flinn, 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, NYU, SITE, Western, and Yale for helpful comments and suggestions. Guan Yi Lin, Hammad Shaikh and Kaili Yang provided 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 seen as an important means of boosting performance – whether for CEOs, sales force personnel, production line workers, or educators.[1] Given that general promise, accountability schemes in an education context have become increasingly widespread, with numerous studies analyzing the impacts of existing programs on educational outcomes.[2] Current research indicates that there may be scope for improving such schemes, and policy makers have a keen interest in doing so, both to raise average performance and address distributional concerns, the latter being especially relevant as differences in outcomes while in school are known to perpetuate, fueling lifelong inequality. Yet from a policy perspective, no existing education research allows policy designers to assess the impacts of different accountability reforms on the entire student performance distribution in a systematic way.

In this paper, we propose an approach to fill the gap. The natural starting place is the relationship between formal incentives and teacher effort. We show this relationship can be recovered when applying a semi-parametric method we develop, leveraging exogenous incentive variation and rich administrative data. Key to the method is the following idea: while effort is typically unobserved, it should (as a productive input) be reflected in observed output. Further, effort should – as is the standard economist's prior – be responsive to changes in formal incentives in a systematic way. Under minimal structure on the technology, unobserved effort can be backed out from observed changes in output, yielding effort as a function of incentive strength.[3] In line with the well-understood incentives implied by threshold schemes, the resulting hump-shaped function peaks where incentives are strongest and declines on either side, although in an asymmetric fashion. Supporting evidence indicates that this pattern is due to teachers adjusting discretionary effort, rather than other relevant inputs being altered (notably, class sizes or teacher and student classroom assignments).

We recover the primitives underlying this effort function by estimating a flexible effort-choice model. The model makes explicit how teacher effort depends on features of the incentive scheme. The model's cost and benefit parameters can be credibly identified, as we show, and the resulting model estimates fit the data closely, including features of the data not targeted in the estimation routine. They indicate that the marginal cost of effort is increasing in the effort devoted to other students in the classroom and that teachers responded to the introduction of NCLB proficiency by boosting effort on a targeted basis.

[1] A vast literature in economics, discussed below, has studied such schemes; Lazear (2000) is a classic example.

[2] Figlio and Loeb (2011) provide an informative survey.

[3] Incentive strength can be captured by the distance between a student's predicted score and the target imposed when the incentive scheme is introduced.

The centerpiece of the analysis is a simulation framework for conducting policy-oriented counterfactuals based on the model and these estimates. The goal is to shed light on the relative merits of alternative incentive systems in education in terms of the *full* distribution of outcomes – the distinctive new feature of the output we produce. Our counterfactual policy framework allows us to explore the relative performance of existing schemes throughout the distribution, and to project the effects of schemes yet to be enacted in a systematic way. In doing so, we can analyze the setting of alternative rewards and targets – key issues in incentive design – while effort adjusts endogenously to the new incentives. The resulting output provides a menu of options enabling policymakers to select their favored scheme according to their preferences over average effort and student outcome inequality.[4]

Three new findings emerge from the counterfactual analysis. First, widely-used 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. We are able to establish this regularity by considering the full range of possible targets in turn, showing that as the fixed target moves up the predicted test score distribution, so average effort increases at the expense of creating a wider outcome distribution. This makes target setting consequential, depending on policy makers' preferences over effort and inequality, our framework offering the first evidence to guide policy makers in terms of how steep the underlying tradeoff is.

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 SD, reduces test score variance by 7.8 percent, and reduces the black-white test score gap by 7 percent – all at no extra cost. Third, student-specific *targets* allow policymakers to reduce inequality without sacrificing average effort. We show that switching from fixed to value-added targets (which are student-specific, the target being a function of the student's prior performance) reduces inequality in teacher effort across students by as much as 90 percent, as value-added targets provide incentives to devote similar effort levels to all students. Perhaps surprisingly, if policy makers place a high priority on limiting outcome inequality, fixed targets can still dominate value-added schemes when the targets are set relatively low in the performance distribution – below the 30th percentile in our analysis.

Overall, the findings draw attention to the scope for improved policy design by applying our approach. We show that feasible schemes yet to be implemented are capable of reducing test score inequality while improving average student performance at no additional cost. Further, by allowing policy makers to gain insight into the distributional consequences of education accountability

[4] This positive emphasis contrasts with the normative approach in the optimal contracting literature; see Mirrlees (1975) and related theoretical studies.

systems, the approach enhances the prospects for using education reforms to combat inequality in a cost effective manner, an especially critical public policy objective today.

The rest of the paper is organized as follows: The next section relates our analysis to the existing literature. Section III describes the incentive variation and the administrative data used in the analysis. Section IV presents the semi-parametric method for uncovering the effort-incentive strength relationship, along with our estimates of the effort function. Section V develops a model of effort setting that can rationalize the estimated function, with Section VI discussing the estimation and identification of the model parameters, and Section VII presenting estimates and model fit. Section VIII sets out our counterfactual framework, Section IX describes the counterfactual results, and Section X concludes.

II. Related Research

This paper builds on several prior literatures, starting with a prominent line of research that studies the introduction of actual incentives in the workplace. Lazear's classic 2000 paper shows how replacing a fixed wage contract with a new piece-rate style incentive scheme by Safelite Glass Corporation led to an increase in company profits. It also draws attention to distributional effects across workers, with high-productivity workers in particular gaining from the new scheme.[5] We develop the inequality theme, considering the implications of incentive schemes for the distribution of *student*, rather than worker, outcomes. The influential study by Bandiera, Barankay and Rasul (2005) also demonstrates that changes in workplace incentives generated significant productivity gains, this time among fruit pickers when moving from a relative incentive scheme to a piece rate. They provide convincing evidence that workers internalize the effects of their behavior on co-workers, a conclusion based on a novel calibration procedure for recovering the parameters that influence worker effort choices. In our study, we are also interested in the parameters governing effort choices when incentives change, and propose an estimation approach using the semi-parametric effort function as an input.[6]

Public education, the context for our study, provides a high-profile policy arena in which incentive schemes have been adopted widely. Given their general aim of increasing teacher and school effort and boosting measured performance, a substantial body of empirical research has already examined the effects of education accountability schemes on student achievement – see

[5] Related to this, Bandiera, Barankay and Rasul (2007) explore how managerial incentives affect the mean *and* dispersion of worker productivity using an experiment that introduced a performance bonus for managers.

[6] Other papers in the literature consider incentive variation more broadly, including Mas and Moretti's (2009) study of the productivity effects of varying peers among supermarket checkout staff, and Bandiera *et al.* (2010), who consider social incentives based on friendship networks in the workplace as an alternative to monetary rewards.

Carnoy and Loeb (2002), Figlio and Winicki (2005), Hanushek and Raymond (2005), Lavy (2009), Dee and Jacob (2011), and Imberman and Lovenheim (2015), among others. Several convincing papers document the way in which proficiency-count incentives have led educators to focus on some groups of students at the expense of others – whether exempting disadvantaged students as in Cullen and Reback (2006) and Figlio and Getzler (2006), or concentrating on students close to proficiency targets rather than students far below or above – see Burgess, Propper, Slater and Wilson (2005), Reback (2008), and Neal and Schanzenbach (2010), for example. Similarly, Deming, Cohodes, Jennings and Jencks (2016) show that schools at risk of being classified as "low performing" under the 1990s accountability program in Texas responded by concentrating effort on lower-scoring students, reflected both in achievement and long-run outcomes. Such varieties of non-uniform attention may be an especial concern when it is disadvantaged students who are neglected. Building on this evidence from existing programs, we study the effects of alternative accountability incentives on the full distribution of student outcomes while in school, including the effects of schemes yet to be enacted.

The semi-parametric approach we propose for recovering the effort function uses an incentive strength measure building on measures in prior work. Two somewhat subtle aspects of our chosen measure are worth highlighting, as they turn out to be important in developing the approach. First, our measure is *continuous* and can be computed for each student.[7] This will allow us to estimate effort at all points in the incentive strength distribution, important for the subsequent estimation and policy analysis. Second, the predicted student scores we use to form the measure are based on *pre-reform* data, enabling us to make plausible baseline effort predictions, as described in Section IV.[8]

The estimable model of teacher effort setting we develop shares several features with a literature that estimates principal-agent models directly using personnel data.[9] In that literature, unobserved effort decisions are cast in terms of an optimal effort choice model given prevailing incentives (as in our analysis); also, the distinctive pattern of output is related to prevailing incentives to infer how optimal (unobserved) effort must be set; and the model is taken to the data to estimate benefit and costs parameters governing worker decisions. In an education context, this type of model-focused approach is rarely used, yet our model is necessary given the main goal of the analysis: to trace the impact of alternative accountability incentives (and the targets and rewards

[7] In related approaches, Deming *et al.* (2016) aggregate incentive strength to the school level, and Neal and Schanzenbach (2010) group students into deciles of the ability distribution.

[8] We calculate expected outcomes using a prediction algorithm similar to Reback (2008) and Deming *et al.* (2016), although those studies do not have access to a pre-reform period.

[9] See Prendergast (1999) for an illuminating survey of the personnel literature. More recently, Copeland and Monnet (2009) provide a sophisticated dynamic analysis of individual worker effort choices in the context of threshold incentive schemes in the check-clearing industry, along with estimates of the welfare costs of higher effort.

or penalties they entail) on the resulting distribution of educational outcomes. Such design issues come naturally to mind when considering the various existing education accountability schemes, including proficiency schemes (such as NCLB) that set fixed performance targets based on school sociodemographics and value-added schemes whose targets condition on prior student scores.[10]

Beyond the impact of existing reforms, incentive designers often wonder about more speculative considerations, looking to the effects of changing the parameters of existing schemes counterfactually, or the effects of incentive schemes yet to be implemented in practice. Approaches that combine a strategy for identifying effort under prevailing incentive provisions with a framework for counterfactual analysis are thus appealing, as in recent research studying worker incentives – see Misra and Nair (2011) for instance.

In this vein, the counterfactual policy framework combining the parameter estimates with the model brings several advantages: It provides a feasible means of constructing alternative incentive schemes that can be fed into the policy analysis. Then it also allows their impacts to be measured on a comparable basis by equating costs. Here, their effects on the *full distribution* of relevant outcomes can be traced for the first time.[11]

III. Institutional Setting and Data

Our analysis requires exogenous incentive variation and rich administrative data. The state of North Carolina provides both.

Incentives: On the incentive front, we make use of the introduction of NCLB provisions in the state in the 2002-03 school year, following the passage of the federal No Child Left Behind Act in 2001. NCLB sought to close performance gaps by requiring schools to meet Adequate Yearly Progress ('AYP') targets for students, while imposing penalties for under-performing schools. We focus on the AYP target shared by all students in a given grade, treating that as a reasonable approximation to the prevailing incentives under NCLB.[12] Doing so provides a potent source of across-student variation in teachers' incentives to devote extra effort, as we will demonstrate.[13]

The federal NCLB program was introduced on top of the state's pre-existing school-based

[10] Studies that focus on particular aspects of accountability schemes already in operation include Cullen and Reback (2006), Neal and Schanzenbach (2010), and Macartney (2016).

[11] A recent experimental study by Loyalka *et al.* (2018) uses random assignment of Chinese elementary school teachers to explore the student achievement effects of incentives based on 'pay-for-percentile' (as in Barlevy and Neal 2012). The 'level' scheme they consider does not feature a test score proficiency target; this, along with the assumed tournament structure, serves to minimize any distributional effects that might arise.

[12] Neal and Schanzenbach (2010) take the same stance.

[13] The NCLB legislation was complex, and provides several other viable sources of variation. For example, given that AYP targets were also set for nine student demographic subgroups, prior work (Reback 2008; Deming *et al.* 2016) has used student subgroup membership to identify accountability pressure across students and schools.

accountability system, the ABCs of Public Education, which applied to all schools serving kindergarten through eighth grade starting in the 1996-97 school year. The ABCs assigned a school-grade-specific target gain to each grade (from 3 to 8), and all teachers and the principal received a monetary bonus if their school achieved its overall growth target, based on average school-level gains across all grades. The implied pre-existing incentives under the ABCs to exert effort throughout the distribution in a reasonably uniform way contrast sharply with the non-uniform incentives under NCLB. How to treat possible interactions between the two is unclear, theoretically and from a practical perspective. The latter in particular will concern us when it comes to estimation.

Aside from formal incentives, with the ABCs focusing primarily on student *growth*, this precursor to NCLB also assigned schools 'low-stakes' status labels based on school proficiency rates to allow parents to keep track of performance. Specifically, the program featured three targets at different points in the score distribution,[14] students achieving proficiency status when their test scores met or exceeded the second target. When NCLB was introduced in 2002-03, North Carolina used the second target as the NCLB proficiency standard, although the first and third may in practice have had some continued salience for educators and parents, a possibility we examine below.

Data: In addition to useful incentive variation, North Carolina offers 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 school and (via a standard matching procedure) to a teacher in any given year. Our main performance variables are constructed from individual student test scores, which 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. The care on the part of psychometricians in designing these tests will be important for comparability.

Our sample period runs from school year 1996-97 to 2004-05 and we restrict attention to students in third to fifth grade, where classes are self-contained and the most accurate matching of students to teachers is possible.[15] These restrictions notwithstanding, our sample is very large, with nearly three million student-grade-year observations.

Table 1 provides descriptive statistics for our sample. In the top part of the table, we summarize the standardized test measures developed in North Carolina as part of the ABCs reform,

[14] The first marked the boundaries between 'insufficient' and 'inconsistent' mastery, the second between 'inconsistent' and 'consistent' mastery, and the third between 'consistent' mastery and 'superior' performance.

[15] We follow prior research studying North Carolina (Clotfelter *et al.* 2006, for example) and take the teacher proctoring the corresponding end-of-grade tests to be the classroom teacher during the school year in these grades.

prior to the enactment of NCLB. Here mathematics scores (in levels) are reported separately in the periods before and after 2000-01 – the academic year North Carolina changed the scale used to measure mathematics scores. These test score levels are relevant under NCLB, which requires that each student exceeds a target test score. As the table shows, both mathematics and reading scores increase monotonically across grades, consistent with knowledge being accumulated in those subjects over time. The dataset also provides useful demographic controls, summarized in the bottom portion of Table 1: individual students' race, disability, limited English proficiency, and free lunch eligibility. In aggregate, 39 percent of students are minorities (non-white), 14 percent are learning-disabled, only 3 percent are limited English-proficient, and 42 percent are eligible for free or reduced-price lunch. Around a quarter of students have college-educated parents.

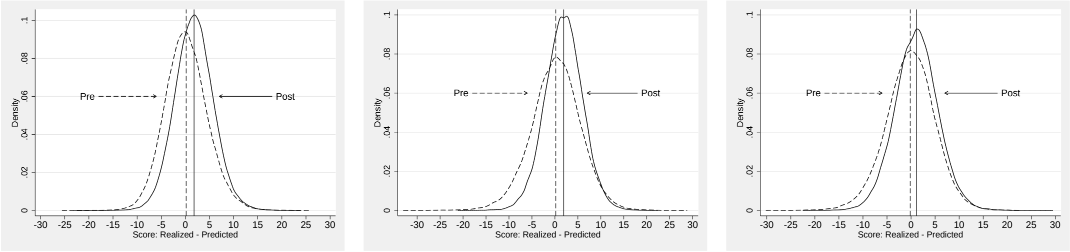
Table 1 – Student-Level Descriptive Statistics

	Full Sample		
	Mean	*SD*	*N*
Performance Measures			
Mathematics Scores			
Pre-2001			
Grade 3	142.87	11.17	396,341
Grade 4	151.56	10.56	384,349
Grade 5	158.18	10.23	376,044
Post-2001			
Grade 3	252.34	7.13	509,571
Grade 4	257.82	8.08	507,622
Grade 5	261.54	9.39	512,425
Reading Scores			
Grade 3	147.03	9.33	901,235
Grade 4	150.65	9.18	887,153
Grade 5	155.79	8.11	883,689
Demographics			
Male	0.51	0.50	2,778,454
Minority	0.39	0.49	2,776,729
Disabled	0.06	0.24	2,778,635
Limited English Proficient	0.03	0.17	2,778,623
Free or Reduced-Price Lunch[(a)]	0.44	0.50	1,998,653
College-Educated Parents	0.25	0.43	2,757,648

Notes: Summary statistics are calculated over all third to fifth grade student-year observations from 1996-97 to 2004-05. [(a)]The free or reduced-price lunch eligibility variable is not available prior to 1998-99.

Initial Evidence of School Responses

We present descriptive evidence indicating clear test score responses to the introduction of NCLB in 2002-03 measured in terms of mathematics scores. The panels of Figure 1 plot the densities of realized test scores in grades three to five in 2002-03 relative to densities of predicted scores in each grade; the latter scores provide a suitable benchmark, being constructed based on all available information about individual students prior to NCLB's introduction, and thus excluding any incentive response to the reform.[16] In each panel, the distribution of the realized test scores shifts right following the introduction of NCLB, consistent with there being an improvement in some unobserved determinant of test scores. A natural interpretation is to view this as an effort response by educators given the heightened incentives – more likely if no other observed inputs changed, an issue we shed light on in what follows.



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

Figure 1 – Densities of Realized Minus Predicted Mathematics Scores (post- versus pre-reform) By Grade

IV. Recovering the Effort Function

In this section, we set out a semi-parametric method for determining the incentive-effort relationship. Then we explain how it is implemented using our administrative data, before presenting the resulting effort function estimates.

IV.A. The Method

Preliminaries. The foundation for our method is a test score technology that relates measured education output y to various inputs. Among those inputs, we place particular emphasis on the discretionary actions of educators that can increase output, given our interest in incentives. We will refer to such actions simply as 'effort.'[17] Effort is not observed directly – a central challenge for researchers studying work incentives. Yet it should, as a productive input, be reflected in observed outputs such as test scores in a systematic way, depending on the incentives in place. This idea serves as the basis for the method we propose.

[16] The next section describes these predictions more fully.

[17] 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."

In our formulation, we will literally take 'effort' to refer to changes in output – observable test scores – that are attributable to incentive variation rather than changes in other inputs, such as teacher quality and class size. We denote the effort directed to student i as e_i, which is endogenous to the prevailing incentive scheme.

Student test scores also depend on various exogenous inputs, such as heterogeneous student ability.[18] Let the exogenous inputs be summarized in a single measure, which we denote $\hat{y}_i$ for student i. Increases in this measure will capture more favorable exogenous 'production' conditions.[19]

Effort and exogenous inputs are assumed to be related in a systematic way via an education production technology to output, which is measured using actual test scores (written y_i for student i). We make the following assumption about this technology.

Assumption 1: The education production function is (a) linear and (b) additive in its inputs (including the production function error).

As we discuss further below, linearity – part (a) – is quite standard in the literature, serving as a reasonable first-order approximation to a richer underlying technology.[20] Additivity of the inputs – part (b) – will be important in order to apply our differencing approach.

Under Assumption 1, we write the education production technology at the student level as

$$y_i = \hat{y}_i + e_i + \epsilon_i, \tag{1}$$

where $\hat{y}_i$ captures exogenous inputs (as noted), e_i is the teacher effort directed to student i, and ϵ_i is a test score shock, representing random determinants unobserved by the econometrician; we place relevant structure on the shock in Assumption 2 below.

Our semi-parametric method utilizes a natural measure of *incentive strength*. This is student-specific, defined as the difference between the student's *ex ante* predicted score ($\hat{y}_i$) and the accountability threshold (y^T).[21] We will label this measure $\pi_i \equiv \hat{y}_i - y^T$, noting that it is *continuous*, and also straightforward to construct for each student in our administrative dataset.

In order to relate this incentive strength measure to effort, we suppose that educators decide systematically how much effort to exert, although the method requires no explicit assumptions about the effort-setting process. The solution to the educator's effort-setting problem, whatever form it takes, will in general yield an *effort function* that maps the underlying production conditions, including characteristics of the student (captured by the student's predicted score) and features of the accountability scheme (the position of the target) to a corresponding effort level. In the

[18] We treat students as passive (rather than active) learners.

[19] We describe how $\hat{y}_i$ is estimated shortly, using exogenous predetermined information.

[20] Our approach generalizes readily to allow for nonlinearities, as we discuss below.

[21] The prediction is *ex ante* in the sense of using all available information *prior* to the accountability reform being implemented when forming in the prediction (and thus excluding incentive-related effort).

environment we have described, those production conditions can be summarized concisely by our incentive strength measure, π_i.

Standard intuition would lead one to expect an inverted-U relationship between incentive strength (as defined) and effort under a threshold scheme.[22] While the precise shape of the effort function $e(\pi_i)$ is unknown in practice, our approach provides a transparent way of uncovering it.

We make the following assumption about the error term:

Assumption 2: Conditional on incentive strength (π_i), the error term in the production technology is mean zero.

This assumption requires that (i) the predicted score is accurate and absorbs all systematic non-effort inputs, and (ii) no other input adjusts systematically with π following NCLB's introduction. We will provide further supporting evidence below, showing that important non-effort inputs remain essentially unchanged following NCLB's introduction.

Our differencing approach involves a post- versus pre-reform contrast. Drawing on the technology specified in equation (1), consider a reform that changes the incentives to exert effort according to a new teacher performance target y^T, and let $\hat{y}_{i,post}$ represent the predicted score in this post-reform environment *excluding* any effort response to the reform. Here, we predict the test score student i would have earned based on pre-reform relationships between output and non-effort inputs, combined with i's characteristics in the post-reform environment – the precise implementation of the prediction procedure is described in the next subsection. We make the following homogeneity assumption about that effort response:

Assumption 3: The effort function e^* is identical for all students with the same incentive strength (π_i).

The post-reform score of student i is then given by:

$$y_{i,post}(\pi_i) = \hat{y}_{i,post} + e^*(\pi_i) + \epsilon_{i,post}(\pi_i)\,, \tag{2}$$

based on equation (1).

The Differencing Method. Based on these elements, the optimal effort response as a function of incentive strength, $e^*(\pi_i)$, can be obtained using straightforward differencing. Start with a given value of the incentive strength measure, π_i. In the post-reform environment, use equation (2) to

[22] Given the incentives under such schemes, the reasoning is made clear in Neal and Schanzenbach (2010), for example. That is, because exerting effort is costly, educators have an incentive to focus on marginal students – those predicted to score close to the passing threshold (whether just below or above) – giving less attention to students predicted to be far below or above. In turn, as a productive input, effort should be reflected in observed output (test scores), peaking close to the threshold and declining on either side of that.

average across all students with the same π,[23] thereby eliminating the i subscript, and move the test score components to the LHS. Doing so gives

$$y_{post}(\pi) - \hat{y}_{post} = e^{*}(\pi) + \epsilon_{post}(\pi)\,, \tag{3}$$

where the RHS contains the effort response as well as any noise in the prediction (with noise being separable, given our technology assumption).

Under Assumption 2, $E[\epsilon_{post}(\pi)] = 0$. Thus the RHS of equation (3) is given by $e^{*}(\pi)$, the desired function. For a given value of incentive strength, unobserved effort is then equal to the LHS components of (3), which are observable or can be computed from observables for a given π.

Discussion. The fact that effort is not observed in most datasets presents challenges for applied researchers who study incentives. Our method allows us to recover effort based on a simple difference between actual and predicted output, constructed appropriately. Further, it yields the complete mapping between effort and incentive strength throughout the distribution of the latter.

We noted that the method does not require any assumptions to be made about the effort-setting process. It does require assumptions about the technology, as contained in Assumption 1. The linearity assumption – part (a) – is typically made in the education and personnel literatures, and serves as a first-order approximation to the unknown true production technology. It can be relaxed by including further interactions among inputs. Additivity – part (b) – is needed in order to construct the relevant difference.

Assumption 2 is plausible in our application, given richness of our individual data and ability to predict observed scores, as evidenced in the pre-reform environment. Assumption 3 can be relaxed to incorporate further heterogeneity – interactions among individual inputs, for example. What we have constitutes a reasonable first pass.

IV.B. Implementation

Implementing the method just described using our administrative data involves the following three steps:

1. Predict student performance using pre-reform data, saving the coefficients from a regression of 2001-02 scores (prior to 2002-03, the year in which the reform came into effect) on cubics in prior 2000-01 mathematics and reading scores, as well as student covariates.[24]

[23] As incentive strength is continuous, in practice we will create bins, described shortly.

[24] The student covariates consist of indicators for parental education, gender, race, free or reduced-price lunch eligibility, and limited English proficiency.

2. Construct the *ex ante* incentive strength measure π_i in the following way:
   (a) predict student performance in 2002-03 ($\hat{y}_{i,post}$), which is the score predicted in the absence of any NCLB effort response, using the saved coefficients from the first step, along with student covariates in 2002-03 and prior test scores from 2001-02;
   (b) using the known NCLB target y^T, compute π_i as the difference between the predicted value $\hat{y}_{i,post}$ (which does not include *additional* effort in 2002-03) and the target (given $\pi_i = \hat{y}_{i,post} - y^T$).

3. Estimate the effort response following the reform's introduction for each value of the continuous incentive strength measure (π_i) as follows:
   (a) take the difference between the realized and predicted scores in the post-reform period ($y_{i,post} - \hat{y}_{i,post}$);
   (b) average this difference across all students within small intervals (bins) of incentive strength, defined by the value of π at the bin's midpoint.[25]

   This step gives the components on the RHS of the following expression, given the observables on the LHS: $y_{post}(\pi) - \hat{y}_{post} = e^*(\pi) + \epsilon_{post}(\pi)$.

Under Assumption 2 (as noted in the explanation of the method), averaging implies the RHS of this last equation isolates $e^*(\pi)$, the desired function. In turn, that function can be obtained via the differencing on the LHS using our administrative data.

IV.C. The Estimated Effort Function

Next we apply the semi-parametric method in our North Carolina context using fourth grade scores, and present the resulting effort profile.[26] Given the incentives under proficiency-based accountability schemes, the largest gains should be apparent for students predicted to score near the proficiency threshold. Figure 2 shows that these are indeed the patterns we find across the predicted test score distribution.

In marked contrast, there is virtually no relationship in the pre-NCLB period between a student's predicted position relative to the proficiency threshold and the gain she experiences over the predicted score on average. (This is entirely as one would expect, given there was no incentive for educators to focus on proficiency prior to NCLB's introduction.)

[25] Formally, for bins of size h, compute $[y_{post} - \hat{y}_{post}](\pi) = \frac{1}{N_\pi} \sum_{i:|\pi_i|<h/2} [y_{i,post} - \hat{y}_{i,post}]$, where $N_\pi \equiv \sum_{i:|\pi_i|<h/2} \mathbf{1}$.

[26] We focus on fourth grade noting that one cannot estimate the pre-reform coefficients for fifth grade given the developmental scale change in 2000-01, and applying our procedure using third grade scores would be likely to conflate the effects of NCLB with a student accountability reform introduced in that grade the previous year.

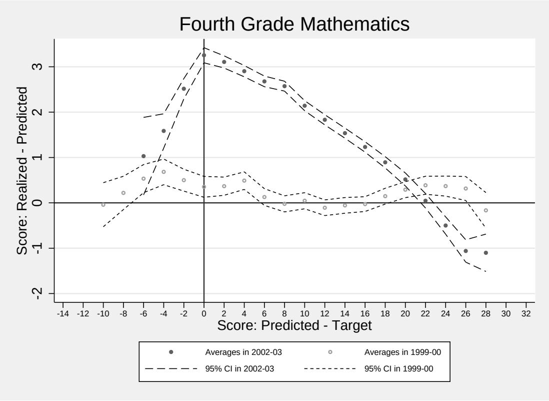


Figure 2 – Effort Responses

The most plausible explanation for semi-parametric inverted-U pattern in Figure 2 involves teachers adjusting discretionary effort in response to incentives. This targeting of 'effort' to marginal students could take a variety of (unobserved) forms: teachers raising their energy levels and delivering material more efficiently to marginal students, or teaching some students more intensively 'to the test.' While we do not have direct information about different actions taken by educators in the classroom, the relatively flat profile pre-reform suggests that there were no persistent factors affecting test scores working at any particular point in the ex ante incentive strength distribution. Then, the post-reform profile takes on a shape that can be rationalized, based on the known incentives.

We are able to examine whether changes in important non-effort inputs to education production might account for the same pattern. As we show in Appendix A.2, the inverted-U pattern is unchanged when controlling for teacher ability or class size, so the incentive effect cannot be explained by school principals assigning marginal students to higher-ability teachers or smaller classes; nor do we find any indication that school principals responded to the incentive reform by making classrooms more homogeneous (in terms of prior test scores), which could otherwise allow for teachers to better target instruction to marginal students without increasing overall effort. We also demonstrate that variation in the incentive response occurs almost *entirely* within-classroom – see Appendix A.3. It is unchanged when conditioning on classroom fixed effects, supporting the view that the incentive response is due to teacher actions. In sum, the available evidence supports our *teacher effort* interpretation.

V. Rationalizing the Effort Function: An Estimable Model

The effort function recovered using our semi-parametric procedure in hand, we now wish to rationalize that function in terms of underlying primitives of the effort-setting process, including those under the control of the policy maker. Doing so is key for the counterfactual analysis of incentive schemes that follows. Our approach will be to develop a flexible effort-setting model, in which teachers choose effort optimally by balancing the benefits against the costs. This provides a natural lens for understanding the semi-parametric pattern.

V.A. The Environment

The general setting. We start by describing features of the environment that public schools typically operate in, to motivate our modeling choices and clarify the likely impact of the types of incentive scheme in education we consider. The general environment is already well-described in Neal and Schanzenbach (2010): we draw on parts of their discussion here.

Public schools are expected to provide a minimum quality level pre-reform, which can be captured by some minimal set of effort levels, and accounting for the financial resources they use; otherwise, severe sanctions would be imposed. At the same time, in the pre-reform environment, schools and their personnel are assumed to enjoy rents. The existence of such rents prompts interest in education reform, intended to raise quality via boosting effort, with various candidate reforms coming to mind. Likely reforms will tend to leave some rents 'on the table' in order to be feasible, although these rents should diminish in line with average effort increasing (as the broad evidence from the accountability literature indicates); thus, for the most part, the participation constraints of administrators and teachers are still satisfied post-reform.[27]

In terms of their objectives, public schools are complex organizations, facing many real constraints unobservable to the researcher. Their production technologies are similarly complex and, as yet, not fully understood. School education production has obvious *team* elements. We will assume that a given school operates as a single decision-making unit, in the absence of a plausible model of the school's internal workings. For our purposes, this means that we will treat the principal and all the teachers as acting co-operatively in order to meet the school objectives. Thus we abstract from any consideration of free-riding on the efforts of other teachers, even though that is a possible issue in a team-production setting, such as a public school, and worthy of study in future. Our implicit assumption is simply that the school is well-managed on a teacher-by-teacher basis and collectively.

The types of accountability reform we will consider are all *school-level* reforms, given they are the predominant mode to date.[28] As we will discuss below, we assume that the principal devolves effort decisions to individual teachers, who are likely to be more familiar with local production conditions. Thus the effort decision under a given accountability reform will be analyzed at the

[27] Following the introduction of NCLB in North Carolina, for instance, we do not see significant outflows of teachers. We will therefore abstract from that margin.

[28] Teacher-level incentive schemes, as in Lavy (2009), are beyond the scope of our analysis.

teacher level. In our setting, this turns out to be a useful approximation: much of the response to NCLB, for example, occurs within-classroom. That this should be so is not obvious *a priori*, but the evidence we discuss supports the teacher-level formulation. Complex aggregation issues across teachers (relating to the attainment of school-level targets, for instance) will be abstracted from.

Model specifics. Having laid out the general context, the model we present has four main elements: a production technology, an accountability incentive scheme, a corresponding expected benefit function, and a cost of effort function. From these necessary elements, we construct the teacher's objective – we motivate the *teacher* focus further below, building on the discussion of the general setting. That objective will allow us to express the teacher's optimal effort choice as a function of key parameters, including features of the incentive scheme. Those parameters can then be estimated on the basis of an estimation routine presented in the next section.

Production Technology: We assume a standard linear and additive structure for the student-level education production technology. In line with the previous section, the technology is given by equation (1):

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

where $\hat{y}_i$ captures all the exogenous inputs for student i, e_i is the teacher effort directed to student i, and ϵ_i is a shock to test scores.[29] We assume that the error has a cumulative density function given by $F(\cdot)$, with mean 0 and variance σ^2.

Accountability Incentives: Accountability schemes define a clear performance metric along with corresponding rewards or punishments. Accordingly, we characterize an *incentive scheme* l by a target level y_l^T and a corresponding reward b_l, both of which are exogenously given. The target could (variously) be an exogenously fixed score, a function of average student characteristics including past performance, or even be student-specific – different possibilities are allowed. The reward parameter b_l governs how target attainment maps into the educator's payoff. This will typically have a formal component – monetary rewards or non-monetary punishments under standard accountability schemes; it may also involve informal components – psychological pressure, for instance. We assume that a teacher receives a benefit b_l for each student in her class whose score exceeds y_l^T (and so is proficient at level l), the benefit combining all formal and informal components. In the absence of further information, we suppose all teachers share the same benefit from attaining target l.

In the context of our study, the NCLB incentive consists of one explicit fixed target level

[29] We omit time subscripts, as the model is estimated using the impact of the NCLB incentive reform in 2002-03, rather than also relying on variation from subsequent years. At least as a first pass, we do not view dynamic considerations as being first order when exploring the distributional implications of rival incentive schemes – the main focus of the analysis.

y_M^T and a reward b_M (the M subscript standing for 'middle'). Because they may be salient for educators and parents, we also allow for the possibility that teachers respond to fixed targets designating other levels of student proficiency (based on the regime in place prior to NCLB), specifically incorporating the high target, denoted y_H^T (where $y_M^T < y_H^T$) with corresponding reward b_H in the objective below.[30]

Benefit of Exerting Effort: Taking the production technology and accountability incentives as given, the teacher assigned to classroom c ('teacher c' for short) derives an expected benefit from exerting effort depending on how that effort affects the probability of a student exceeding each of the targets, summed across all students in her class – as we will discuss, the evidence supports this teacher focus. At a general level, we will write this benefit as $B(e_1, \ldots, e_{N_c})$.

In practice, it is possible that teachers seek to overshoot the prevailing targets in order to protect against potential negative shocks to test scores. We allow for this possibility in a straightforward way: for each target y_l^T, overshooting is represented by a shift parameter, d_l, which moves the effective target.[31] Note that we suppress teacher and school heterogeneity in the shifter parameters. This is not simply for tractability: we provide clear evidence indicating that the semi-parametric effort function is invariant to teacher ability, class size, and the prior preparation of a teacher's students relative to the school distribution (see Appendix A.3). The benefit is then given by:

$$\begin{aligned} B(e_1, \ldots, e_{N_c}) &= b_M \sum_{i=1}^{N_c} Pr[y_i > y_M^T + d_M] + b_H \sum_{i=1}^{N_c} Pr[y_i > y_H^T + d_H] \\ &= b_M \sum_{i=1}^{N_c} \Big[1 - F(y_M^T + d_M - \hat{y}_i - e_i) \Big] \\ &\quad + b_H \sum_{i=1}^{N_c} \Big[1 - F(y_H^T + d_H - \hat{y}_i - e_i) \Big] , \end{aligned} \tag{4}$$

where N_c is the number of students in the class taught by teacher c.

Cost of Exerting Effort: Teacher c faces a cost that is convex in effort applied to each student.

[30] In Appendix C, we rule out a simpler model that only allows for teachers to respond to the NCLB target, as it departs substantially from the semi-parametric effort profile estimated in the previous section. That profile will be used to discipline the model. On that same basis, we also reject three other alternative models, allowing (respectively) for a form of complementarity in production, peer spillovers, and teacher error in the prediction of student ability. As we will show, our proposed model is appealing in that it fits the data very well while also being reasonably parsimonious.

[31] Such overshooting offers an additional degree of freedom when matching the effort profile. Indeed, it is necessary to rationalize maximal effort being directed toward students with predicted scores that equal the target (consistent with the effort pattern recovered in Section IV.C).

We assume the cost function has the following flexible form:

$$C(e_1, \ldots, e_{N_c}) = \frac{\psi}{2}\left[\sum_{i=1}^{N_c} e_i^2 + \ \theta\left(\sum_{j=1}^{N_c} e_j\right)^2\right]. \tag{5}$$

The ψ parameter allows the marginal cost of effort to be scaled, while the parameter θ governs the extent to which effort choices across students in a given class are interdependent.[32] In the case where $\theta = 0$, the cost side collapses to one in which education provision amounts to fully individualized tutoring, while $\theta > 0$ allows the incremental cost of raising the effort supplied to a given student to be higher the more energy the teacher supplies to the rest of the class. This parameterization will enable us to assess whether such a spillover component is important in practice.

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 challenging – this is in contrast to a firm setting, where profit maximization is often a reasonable approximation. In our application, we leverage the fact that an explicit portion of the objective is known as a consequence of a formal accountability scheme being in place.

Taking the above elements together, we can write down the educator objective under different incentive schemes – necessary for exploring the implications of counterfactual targets. We focus on the effort decisions of teacher c, allowing each student to receive student-specific effort e_i. This is reasonable given clear evidence that the teacher effort response is driven almost entirely by within- rather than across-classroom variation in production conditions (summarized by $\hat{y}$) (see Appendix A.3). Because formal incentives under NCLB apply to schools as a whole, it may seem surprising that effort decisions would be taken by individual teachers, in line with that evidence; a natural interpretation is that school principals delegate 'local' decisions to teachers in service of school-level objectives, given that teachers are likely to know each of their students well (and better than the school administration).[33]

Given the teacher focus, teacher c thus chooses a set of effort levels $\{e_1, \ldots, e_{N_c}\}$ to maximize the difference between her expected benefit of effort and the total effort cost:

$$U = B(e_1, \ldots, e_{N_c}) - C(e_1, \ldots, e_{N_c}), \tag{6}$$

[32] The parameter ψ is not separately identified from b_M or b_H, as we show in Section VI.B.

[33] If well-managed, it is likely the entire school would agree how to respond to NCLB, but individual teachers would be left to manage their classrooms by determining how best to apply their effort, keeping the agreed-upon overall objective in sight. While this does not preclude the school administration (the principal, for instance) from also taking actions that are observationally equivalent to teacher effort, we rule out the most obvious such actions as drivers of our results in Appendix A.2.

where the explicit functions are given by equations (4) and (5).

V.B. Optimal Effort

For the given test score targets, $\mathbf{y^T} = \{y_M^T, y_H^T\}$, and predicted scores for the relevant class, $\{\hat{y}_i\}_i^{N_c}$, optimal effort for every student taught by teacher c is jointly determined from the N_c first-order conditions obtained by maximizing equation (6) with respect to the effort each student receives from the teacher. The first-order condition for the effort teacher c directs toward student i is given by:

$$\begin{aligned} \frac{b_M}{\psi} f\big(y_M^T + d_M - \hat{y}_i - e_i^*\big) \ &+ \frac{b_H}{\psi} f\big(y_H^T + d_H - \hat{y}_i - e_i^*\big) \\ &= \left[e_i^* + \theta \sum_{j=1}^{N_c} e_j^* \right], \ \forall\ i = 1, \ldots, N_c\,, \end{aligned} \tag{7}$$

where the first row gives the marginal benefit and the second row, the marginal cost. The optimal effort that solves this equation can be expressed as a function of (i) the model's parameters, (ii) the incentive targets, (iii) the student's predicted score $\hat{y}_i$, and (iv) classroom factors. The latter refer to the classroom-specific distribution of predicted scores $\hat{\mathbf{y}}_c \equiv \{\hat{y}_j\}_{j \neq i}$.[34]

The Solution – Intuition: As specified, optimal effort e^* does not have a closed-form solution.[35] Still, the model structure allows us to provide intuition for the way optimal effort is determined, especially how e^* will change depending on student incentive strength. Here we place additional structure on the form of the benefit, assuming that the density of the test score shock is unimodal – we impose normality on the error for convenience.[36]

To begin with, suppose there is just one target. The first-order condition for the effort directed toward student i then simplifies to:

$$\frac{b_M}{\psi} f\big(d_M - \pi_i - e_i^*\big) \ = \left[e_i^* + \theta \sum_{j=1}^{N_c} e_j^* \right], \tag{8}$$

using equation (7) and substituting our measure of incentive strength $\pi_i \equiv \hat{y}_i - y_M^T$. The marginal benefit of effort will take on the bell shape associated with the assumed normal density of the test

[34] Recall that a non-zero value of the cost parameter θ implies that the effort applied to one student in the class depends on the effort devoted to all other classmates. If that is the case, then the optimal level of teacher effort directed to a particular student should depend on her place within the classroom distribution of predicted scores. As a result, two otherwise identical students who face different classroom distributions may receive different levels of effort – hence the relevant conditioning variable.

[35] The derivatives of the parameters in the nonlinear first-order condition depend both on the effort level of interest and the parameters themselves, precluding a closed-form solution. Instead, an iterative process is required to determine optimal effort.

[36] Looking ahead to the estimation section, this will be in line with Assumption 4.

score noise, scaled up or down by $\frac{b_M}{\psi}$. The peak of the marginal benefit curve occurs at effort level $e_i^{peak} \equiv d_M - \pi_i$, defined as the effort for which the argument of $f(\cdot)$ is zero: for large negative and positive values, the marginal benefit is lower. As π_i (and thus $\hat{y}_i$, given the definition of π_i) increases, it follows that the marginal benefit curve shifts leftward.

First, consider the case where θ is set to zero. The marginal cost of effort is a straight line through the origin with a slope of one. Given the unimodality of the density function $f(\cdot)$, teacher effort will follow an *inverted-U profile* as a function of π, consistent with the empirical patterns documented in Section IV.C. Further, the asymmetry we find in the effort function is a natural consequence of the first-order condition equating the marginal benefit with the upward-sloping marginal cost, where the marginal benefit curve has the assumed unimodal shape.

Next, suppose θ is non-zero. In this more general case, the same type of inverted-U pattern emerges. The main difference is that aggregating the cost of effort to the classroom level allows for 'negative' values of effort in our model (interpreted as students realizing lower gains over predicted scores than in the pre-reform period), which affords a better match with the effort profile. To see how negative values of effort are allowed, note that the marginal cost of effort for each student i becomes a straight line with slope $(1+\theta)$ and vertical intercept $\theta \sum_{j \neq i}^{N_c} e_j^*$.[37] When $\theta > 0$ and average classroom effort is positive, the model permits solutions to equation (8) in which optimal effort is negative.

Now consider the case in which the higher target also influences effort setting, alongside the NCLB target. As reflected in equation (7), the optimal effort profile will now be the vertical summation of the effort profiles implied by each of the two separate targets. The higher target will give rise to a second peak in the effort profile, and will provide a fruitful way to explain the estimated shape of the effort function, as we will discuss below.

Properties: Building on the discussion of the way teacher effort is determined in our model, we are interested in exploring the relationship between the resulting optimal effort profile, traced across all student types $\{\hat{y}_i\}$, and the model parameters. Write optimal effort for student i (determined as the solution to equation (7)) as $e^*(\beta; \hat{y}_i, \mathbf{y^T}, \hat{\mathbf{y}}_c)$, where $\beta \equiv \{\frac{\mathbf{b}}{\psi}, \mathbf{d}, \theta, \sigma^2\}$. For any given values of the model's parameters, the first-order conditions allow us to recover a set of effort levels (across all students) that maximize the teacher objective, summing across all teachers.[38]

Several comparative static results emerge. First, the *spread* of the effort profile is increasing in σ^2, since the marginal benefit curve broadens as σ^2 increases, slowing the rate at which effort

[37] If θ is very small and $\sum_{j \neq i}^{N_c} e_j^*$ is relatively large – as we will find – then θ only has a first-order effect on the intercept.

[38] Appendix D provides a detailed analysis.

declines away from its peak as it shifts against the marginal cost curve (where the intersection determines the effort solution) – this relationship is depicted numerically in panels (a) through (c) of Figure H.1. Second, the proportion of the optimal effort profile taking negative values is increasing in the spillover parameter θ – a straightforward consequence of the vertical intercept of the marginal cost curve shifting up (and the horizontal intercept shifting left) as θ becomes larger (see Figures **??** and **??**).[39] Third, allowing for more than one target, the horizontal location of the two peaks is determined by d_M and d_H, respectively, with each peak shifting right as the associated 'shift' parameter increases – a property illustrated for the M target in panels (d) through (f) of Figure H.1. This occurs since each parameter affects the associated peak of the marginal benefit curve, which affects the maximum of the optimal effort profile. Fourth, the height of the two peaks is increasing in $\frac{b_M}{\psi}$ and $\frac{b_H}{\psi}$, respectively – a property illustrated for the M target in panels (g) through (i) of Figure H.1. As these parameters multiply the density function, each one affects the height of its peak on the marginal benefit curve, which in turn affects the height of its peak in the optimal effort profile.

These properties will be useful in the discussion that follows.

VI. Model Estimation and Identification

This section describes the estimation of the model and the identification of the model parameters.

VI.A. Estimation

Our estimation strategy addresses two issues: teacher effort is an unobserved input, and optimal effort in the model does not in general have a closed-form solution.

Using the production technology in equation (1) and the first-order condition defining optimal effort in equation (7) implicitly, we can write the test score for any student i as

$$y_i = \hat{y}_i + e^* \left(\beta; \hat{y}_i, \mathbf{y^T}, \hat{\mathbf{y}}_c\right) + \epsilon_i \,. \tag{9}$$

Our estimation routine selects values for the parameters $\beta \equiv \{\frac{\mathbf{b}}{\psi}, \mathbf{d}, \theta, \sigma^2\}$ that maximize the joint likelihood of observing the student test score outcomes in the data. To form the likelihood, we use equation (9) and make a further distributional assumption:

Assumption 4: ϵ_i is normally distributed, with mean 0 and variance σ^2.

We can then write the individual likelihood function for any student i as a function of data on

[39] The reasoning is as follows: The marginal benefit curve can only take on positive values (as it consists of a scaled density function), which means that the marginal cost must also be positive whenever the two curves intersect. Thus, effort can only take on a negative value if the horizontal intercept of the marginal cost curve is itself negative.

observed and predicted scores, as well as the effort level implied by the model's parameters:

$$\begin{aligned} L_i(\beta) &= f\left(\epsilon_i \mid \frac{b_M}{\psi}, \frac{b_H}{\psi}, d_M, d_H, \theta, \sigma^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{1}{2\sigma^2} \cdot \Big(y_i - \hat{y}_i - e^*(\beta; \hat{y}_i, \mathbf{y^T}, \hat{\mathbf{y}}_c)\Big)^2\right\} . \end{aligned} \tag{10}$$

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

$$\begin{aligned} \ell(\beta) &\equiv \sum_{i=1}^{N} \log L_i(\beta) \\ &= -\frac{N}{2} \cdot \log(2\pi) - \frac{N}{2} \cdot \log \sigma^2 - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{N} \Big(y_i - \hat{y}_i - e^*(\beta; \hat{y}_i, \mathbf{y^T}, \hat{\mathbf{y}}_c)\Big)^2 . \end{aligned} \tag{11}$$

The Routine: We use a three-step iterative procedure to estimate the maximum likelihood parameter vector, $\widehat{\beta}$. The first step begins with a guess for the value of the parameter vector and solves equation (7) for $e^*(\beta; \hat{y}_i, \mathbf{y^T}, \hat{\mathbf{y}}_c)$, thus determining model-implied effort for each student.[40] In the realistic case in which $\theta \neq 0$, the effort devoted to each student depends on the effort received by all other students in her class; thus this step involves solving a system of N_c equations for each classroom c. In the second step, we use the optimal effort levels to calculate a model-implied test score for each student as $\hat{y}_i + e^*\left(\beta; \hat{y}_i, \mathbf{y^T}, \hat{\mathbf{y}}_c\right)$, which then allows us to evaluate the likelihood in equation (11) in the third step, the resulting parameter estimates then feeding back into the first step above.[41] The routine continues iterating over possible parameter vectors in this way, stopping once it finds the parameter vector that maximizes the likelihood.

To estimate the model, we use the sample of fourth grade students in 2002-03 with non-missing data for actual mathematics and predicted test scores.[42] The proficiency target is set equal to 247 developmental scale points, the actual NCLB proficiency target (y^T_M) in fourth grade; similarly, the target required for 'superior' performance (y^T_H) is set equal to 258 scale points, the corresponding fourth grade value. We also restrict the sample to students in classrooms with at least 7 and no more than 40 students, given our interest in teachers redistributing effort across students *within* classrooms. In practice, this restriction does not affect our estimates.

Descriptive statistics for the sample used to estimate the model, presented in Table H.1, make clear this sample looks quite similar to the full sample of students (see Table 1) in terms of demographic variables, although mean test scores are slightly higher in the former.

[40] See Appendix E for a description of the numerical approach.

[41] Estimation is carried out in MATLAB using the 'fmincon' package.

[42] The focus on fourth grade follows the justification in the semi-parametric approach above.

VI.B. Identification of Parameters

Before estimating the model, we discuss the identification of each of the main parameters of interest.

The σ^2 parameter: This parameter captures the variability in effort (given by $y_i - \hat{y}_i$ for student i) from the model's predictions. Based on the first-order conditions of the maximum likelihood objective function, the parameter σ^2 is equal to the average of the sum of squared deviations between the estimated effort function and effort implied by the model. This variance equals the true variance of test score shocks, using the production technology. As such, the parameter is identified outside the teacher's problem and does not depend on the values of the other parameters.

The θ parameter: This governs the within-classroom tradeoffs in effort that teachers must make across their students. The evidence in Section V.B indicates that $\theta > 0$, given the estimated effort profile shows positive average classroom effort $\sum_{j \neq i}^{N_c} e_j^*$ along with some 'negative' effort values.[43] A positive θ implies that the marginal cost of effort directed toward any student i is an increasing function of the effort directed toward any other student. Exact identification of θ follows from the values of $\hat{y}$ (on the left and right of the peak) for which the estimated effort profile turns negative, conditional on average classroom effort and σ^2.

The $\frac{\mathbf{b}}{\psi}$ and d parameters: The height of the M and H peaks is influenced by $\frac{b_M}{\psi}$ and $\frac{b_H}{\psi}$, respectively, and the horizontal location of each peak, by d_M and d_H, respectively (as in Section V.B). Suppose that the M and H targets are far enough apart so that they do not interact in determining optimal effort around each peak – in practice, the targets are far apart (with $y_H^T - y_M^T \approx 11$), making any interplay between the response to each target unlikely..[44] This allows us to consider identification of the M- and H-related parameters separately. The coordinates of the peak of the effort profile corresponding to each of the two targets come from the estimated effort profile in Section IV.C. We define effort at the M peak to be $e^*_{peak,M}$ (the vertical coordinate) and the incentive strength at the peak to be $\pi_{peak,M}$ (the horizontal coordinate). We have $e^*_{peak,M} = \frac{b_M}{\psi} f(0) - \theta \sum_{j \neq i}^{N_c} e_j^*$, since peak effort occurs at $d_M = \pi_{peak,M} + e^*_{peak,M}$. Given that $e^*_{peak,M}$, $f(0)$ and $\theta \sum_{j \neq i}^{N_c} e_j^*$ are known quantities, this implies that $\frac{b_M}{\psi}$ is identified.[45] The parameter d_M is then identified from $d_M = \pi_{peak,M} + e^*_{peak,M}$. An analogous argument can then be used to identify $\frac{b_H}{\psi}$ and d_H, using the coordinates for the H peak.

[43] Negative effort values can only arise if the marginal cost curve is shifted upward from the origin; given that the vertical intercept is $\theta \sum_{j \neq i}^{N_c} e_j^*$, this implies that θ and $\sum_{j \neq i}^{N_c} e_j^*$ have the same sign.

[44] We expand the argument to allow for possible interactions in Appendix D.

[45] Here, $f(0) = \frac{1}{\sqrt{2\pi\sigma^2}} \approx 0.1$, given an estimate of 15.7 for σ^2.

VII. Model Estimates

The parameter estimates and evidence of model fit are described next.

VII.A. Estimated Parameter Values

Table 2 presents the estimates of the model's parameters. In terms of the cost side of the model, the estimates indicate that (as expected) it is costly for teachers to exert effort ($\frac{b_M}{\psi} > 0$) 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 ($\theta > 0$).

The estimates also indicate that teachers reacted strongly to NCLB's introduction, in two ways: First, considering the response to the actual NCLB proficiency target (y_M^T), teachers had an incentive to try harder in order to guard against the possibility of a negative test score shock. Such behavior is observationally equivalent to teachers acting (under our formulation) as if the test score proficiency target (y_M^T) were higher than its mandated level. Here, the estimate of $d_M = 3.19$ implies that teachers behave as if the effective target were over 3 developmental scale points higher than the mandated target.

Second, teachers would also be led to exert additional effort if they were responding to the high target (y_H^T), which marks the difference between 'proficient' (required for NCLB) and 'superior' performance. Here, the estimated ratio $\frac{b_H}{\psi}$ is positive and significant. This is consistent with teachers behaving as though there were additional benefits to helping students clear the threshold for superior performance, despite this standard not being legislated by NCLB.[46] It is worth noting that the estimates imply a more muted response to the high target than the proficiency target. Specifically, the estimated benefit of helping students clear it is two-thirds of the benefit of helping them clear the proficiency threshold ($= 24.001/36.297$), and teachers also appear to overshoot the high performance target ($d_H = 1.634$) by around half the overshooting that occurs for the NCLB proficiency target ($d_M = 3.19$).

VII.B. Model Fit

In terms of model fit, we start by plotting the 2002-03 data from Figure 2(a) along with the effort predicted by the model. We use the model to generate effort levels for each student and then collapse model-implied effort levels into binned means along the horizontal axis for visual ease. It is clear from the figure that the model fits the data very well, as its mean effort prediction is within

[46] One rationalization for this is that NCLB made school performance (in terms of attaining targets) more salient, especially to parents, and schools wished to demonstrate that better prepared students also gained following NCLB's introduction.

Table 2 – Parameter Estimates

Parameter	Estimate
$\frac{b_M}{\psi}$	36.297*** (0.6234)
$\frac{b_H}{\psi}$	24.001*** (0.5450)
d_M	3.1945*** (0.0647)
d_H	1.634*** (0.0950)
θ	0.0075*** (0.0010)
σ^2	15.702*** (0.0084)
N	89,271

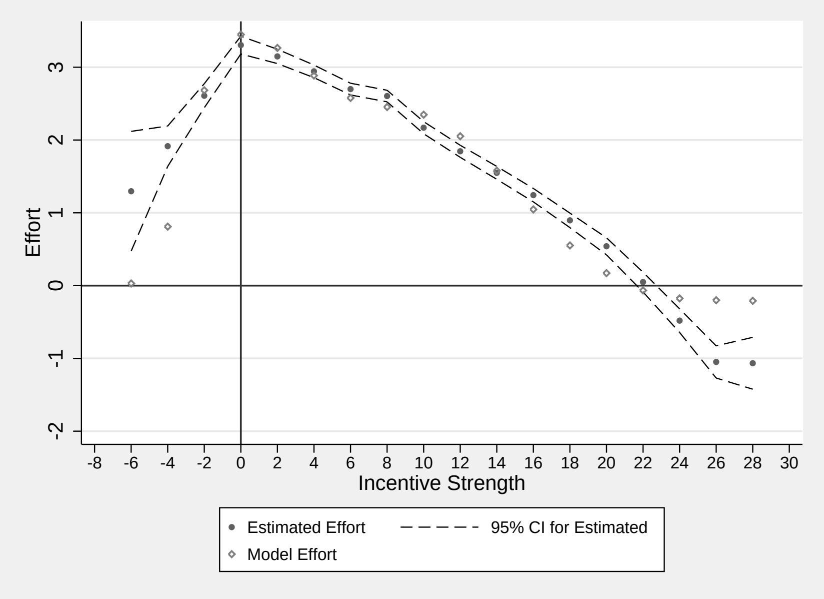
Notes: Standard errors appear in parentheses and are calculated using the outer-product of gradients method. *** denotes significance at the 1% level.

or very close to the confidence intervals of the means from the data in almost all cases, aside from a few points, including those closer to the far right and left of the incentive strength distribution.

Next, we can use equation (9) to predict student test scores based on the model and assess how well the model matches various test score moments. To that end, Table 3 shows comparisons of observed and model-predicted moments across several student subgroups. It is worth noting that our estimation routine does not target these subgroup moments directly, making the comparisons an informative test of the model's fit. All numbers in Table 3 are rounded to 2 decimal places. It is clear that the fit is *exceptionally* tight, with the model and the data usually differing only at the third decimal place (not reported in the table for expositional clarity).

The within-classroom variance of the realized mathematics score accounts for 74 percent of the overall variance across all students (a statistic not reported in the table). Alongside that, the within-classroom variance of the predicted mathematics score from the model accounts for 80 percent of the overall variance, implying that the model replicates the sources of test score variation in the data in an accurate way.

Figures B.1 and B.2 in Appendix B illustrate the model's fit further 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 very closely indeed.



Notes: This figure presents the profile of the estimated effort function in 2002-03 from Figure 2(a) along with the 95 percent confidence intervals for that function and the binned means of model-implied effort.

Figure 3 – Inverted-U Response to NCLB and Model Fit

VIII. Counterfactual Framework

In this section, we set out our counterfactual approach. By combining the plausible structure of the model with the estimates from the previous section, we can explore a variety of alternative proficiency targets and their effects on outcomes in a systematic way; of note, the designation of marginal students adjusts endogenously as incentive provisions change. In turn, we can assess the effects of different accountability schemes on the *entire distribution* of student outcomes, both in terms of the effort students receive and the test scores that result. This includes students who are marginal with respect to the target under proficiency schemes – the main focus of prior research[47] – as well as the remainder, who constitute the majority.

Key to our approach is the simulation framework presented next. We then describe the set of proficiency targets we consider using the framework, and a cost-equating procedure to ensure comparability, before turning to the counterfactual results themselves in the following section.

VIII.A. Framework

Our simulation framework has three elements, each necessary to simulate the full counterfactual score distribution under a given accountability scheme. These are: the incentive parameters of that accountability scheme, an effort-setting condition under the counterfactual incentives, and an

[47] See, for example, Reback (2008), Neal and Schanzenbach (2010), Ladd and Lauen (2010), and Deming *et al.* (2013).

Table 3 – 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 (0.19)	0.96 (0.19)
White	0.98 (0.14)	0.98 (0.13)
Black	0.92 (0.27)	0.92 (0.27)
College-Educated Parents	0.99 (0.11)	0.99 (0.10)
Non-College-Educated Parents	0.94 (0.23)	0.94 (0.23)
Economically Disadvantaged	0.93 (0.25)	0.93 (0.25)
Non-Economically Disadvantaged	0.99 (0.12)	0.99 (0.12)
Mathematics Score		
Overall	259.51 (7.17)	259.47 (7.19)
White	261.49 (6.82)	261.46 (6.84)
Black	255.71 (6.36)	255.70 (6.38)
College-Educated Parents	262.73 (6.70)	262.72 (6.79)
Non-College-Educated Parents	257.26 (6.61)	257.20 (6.57)
Economically Disadvantaged	256.54 (6.51)	256.47 (6.48)
Non-Economically Disadvantaged	261.99 (6.73)	261.97 (6.79)

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

education production technology that incorporates effort, generating the counterfactual test scores as output.

Accountability Schemes: These schemes can each be characterized by a set of targets and bonuses (or punishments). Under NCLB, both the proficiency target y_M^T and bonus payment b_M are taken to be constant across all students, as is appropriate. In this section, we consider several different counterfactual targets beyond those implemented in practice, along with two contrasting weighting systems for implementing differential bonus payments across students. In specifying alternative targets and bonuses, we therefore allow for the possibility that these may be student-specific. Accordingly, we will write the proficiency target (superscripted 'T') for student i at time t as y_{it}^T and the student-specific bonus $b_i \equiv w_i \cdot b_M$, where w_i is a weight placed on student i that allows the bonus, b_M, paid for each proficient student to be scaled heterogeneously.

Effort Setting: Our primary interest is in the way accountability incentives influence teacher effort. In line with the model presented in Section V, we will think of effort as being the result of a *teacher* optimization problem. Specifically, teacher c chooses a set of effort levels in period t, $\{e_{1t}, \ldots, e_{N_c t}\}$, one for each student in her class, to maximize the objective given by a variant of equation (6).[48]

In our simulation framework, we let $\hat{y}_{it}$ be student i's predicted score in year t and continue to define $\hat{\mathbf{y}}_c$ as the classroom-specific distribution of predicted scores. We hold fixed students' predicted scores in the absence of accountability incentives across all of our counterfactual simulations, and keep the model's parameters set at their estimated values ($\widehat{\beta} \equiv [\widehat{\frac{b_M}{\psi}}, \widehat{d}_M, \widehat{\theta}, \widehat{\sigma}^2]$). In each simulation, we either set new proficiency targets (y_{it}^T) or change the bonus paid per proficient student ($b_i \equiv w_i \cdot b_M$) by multiplying the parameter estimate $\widehat{\frac{b_M}{\psi}}$ by a student-specific weight w_i when we wish to make bonus payments vary across students. Taking as given students' predicted scores and the model's underlying parameter values, we then use the updated proficiency targets and bonus payments to compute optimal effort under each counterfactual simulation. Here we follow the same computational procedure as in the model above (described in full in Appendix E): that is, we solve N_c first-order conditions in each classroom simultaneously to recover the full distribution of effort. Optimal effort for student i is then given by $e_{it}^* = e^*\big(w_i, y_{it}^T; \widehat{\beta}, \hat{y}_{it}, \hat{\mathbf{y}}_c\big)$.

Technology and Counterfactual Output: With the counterfactual effort vector in hand, we then use the technology in (9) along with the distribution of test score shocks to obtain the implied test score for any student i under proficiency target y_{it}^T and corresponding bonus payment regime $w_i \cdot b_M$ according to:

$$y_{it} = \hat{y}_{it} + e^*\big(w_i, y_{it}^T; \widehat{\beta}, \hat{y}_{it}, \hat{\mathbf{y}}_c\big) + \epsilon_{it}, \tag{12}$$

where $\hat{y}_{it}$ is the student's predicted score based on all prior information,[49] $e^*\big(w_i, y_{it}^T; \widehat{\beta}, \hat{y}_{it}, \hat{\mathbf{y}}_c\big)$ is the optimal effort level directed to that student under the counterfactual incentive scheme, and ϵ_{it} is an error term that reflects unobserved determinants of the test score. We assume that the test score shock faced by student i is given by ϵ_{it}, distributed as $N(0, \widehat{\sigma}^2)$, where the variance is estimated previously. Equation (12) can then be used to recover the associated test score distribution across all students.

[48] We will set $\frac{b_H}{\psi} = 0$, effectively constraining teachers to respond only to the real NCLB proficiency target.

[49] We described the general prediction approach in Section **??**. Applying that here, $\hat{y}_{it}$ is constructed using a prediction equation that is estimated in pre-NCLB period. It therefore represents an *ex-ante* prediction of each student's test score that does not contain any incentive response.

VIII.B. Counterfactual Incentives

With that basic structure in place, our counterfactual exercises involve setting different student proficiency targets and bonus payments, then exploring the consequences of those for test score outcomes using the estimated production technology. We consider a variety of relevant incentive schemes, including ones that go beyond schemes currently implemented.

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 (with no i-subscript). Payoffs under the fixed scheme are determined by a threshold rule, given by $b_i \cdot 1(y_{it} \geq y^T)$, where b_i 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). The underlying predicted score distribution determines how many students are likely to be in the vicinity of a given target: based on the standard intuition, captured in our model, those marginal students should be expected to receive most effort under fixed target regimes.

Our interest centers on the effects of moving the fixed target through the predicted score distribution in a counterfactual way. Here, the actual NCLB target provides a useful benchmark: we explore the effects of setting targets that differ from this, using the model to determine the associated effort decisions and the implied test score distribution in each counterfactual instance. In total, we cover a range of different settings – seven in total – spanning the full predicted score distribution.[50]

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 affords policymakers an additional degree of freedom with which to improve outcomes. For concreteness, 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; in the second, the weight increases linearly in students' predicted scores, thereby creating incentives to favor higher-performing students (see Appendix F.2).

Value-added ('VA') Schemes: These schemes set targets that are student-specific, depending on a student's prior-year test score, $y_{i,t-1}$. In our formulation, we will express the VA target for student i by $y^T_{it} = \delta + \alpha y_{i,t-1}$, the relevant threshold benefit rule written $b_i \cdot 1(y_{it} \geq y^T_{it})$ (as before). The parameter δ influences the mean of the incentive strength $(\hat{y} - y^T)$ distribution, while

[50] See Appendix F.1.

α governs the variance of that distribution.[51]

We explore the effects of different VA targets on outcomes by varying the target parameters systematically, as follows: each value-added target we consider can be linked precisely to a corresponding fixed target, noting that a fixed target is a special case of a VA target where $\alpha = 0$ and $\delta = y^T$. Taking a given fixed target (the NCLB benchmark score of 247, for example), then for any multiplicative coefficient, α, we choose $\delta(\alpha)$ so that the mean of the resulting incentive strength distribution under the VA target matches the mean under the given fixed target.[52] In the counterfactuals below, for each fixed target we analyze, we consider a host of different values for the multiplicative coefficient, α – twelve in total, in the range 0.1 to 1.9. In doing so, we place more or less emphasis on the prior score, thereby considering the effects of using VA targets to change the spread of the incentive strength distribution (relative to a given fixed target) while leaving the mean unchanged.[53]

VIII.C. Cost Equating

We place all the counterfactual incentive schemes under consideration on a common footing for comparability. To that end, we ensure that every target regime results in the same cost, changing the bonus payment until we achieve cost-equivalence across regimes. Under a constant bonus scheme, equating costs across schemes is equivalent to preserving a given statewide proficiency rate, recalling that the state must pay a bonus for each student deemed proficient.[54]

The essence of the cost-equating procedure is as follows:[55] Teachers' optimal effort choices are influenced by the parameter $\frac{b_M}{\psi}$, which appears in the effort-setting first-order condition. While we cannot separately identify the bonus payment b_M in our estimation framework, we normalize b_M to one and multiply the estimate of $\frac{b_M}{\psi}$ by a constant k: setting $k < 1$ is equivalent to decreasing the bonus payment and setting $k > 1$, to increasing it. Under each target regime, we pick the value of k that equates the cost to the actual cost (equivalent to the passing rate) under the prevailing 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.

[51] To see why, note that the mean VA target across all students is given by $\bar{y}_t^T = \delta + \alpha \bar{y}_{t-1}$ and the variance is given by $var(y_t^T) = \alpha^2 var(y_{t-1})$. Therefore, one can shift the mean by varying δ and manipulate the variance by changing α.

[52] Thus under the NCLB benchmark, for instance, setting $\delta(\alpha) = 247 - \alpha \bar{y}_{t-1}$ implies that the mean of the VA targets (across all students) is 247. It follows that the mean of incentive strength – or $(\hat{y} - y^T)$ – under both the fixed and VA targets is $\bar{\hat{y}}_t - \bar{y}_t^T = \bar{\hat{y}}_t - 247$, where $\bar{\hat{y}}_t$ is the mean predicted score.

[53] Appendix F.3 describes the construction of VA targets in detail.

[54] The target cost that all regimes are equated to involves a proficiency rate of 0.96, the observed rate in 2002-03 under the actual NCLB target.

[55] Appendix F.4 gives a fuller description, including the cost-equating procedure when bonuses are student-specific.

IX. Counterfactual Results

This section presents the main results of our counterfactual analyses. We first consider the outcome distributions associated with fixed targets when bonus payments are the same across all students. Then we show how heterogeneous bonus payments further influence outcomes under fixed targets, before documenting the outcomes under value-added targets (alongside fixed targets that are directly comparable). Because we recover the full counterfactual test score distribution in each instance, we are able to compute a variety of informative 'output' measures. In what follows, we will focus specifically on mean effort and the dispersion of test score outcomes, although other measures are easily computed.

IX.A. Fixed Targets with Homogeneous Bonus Payments

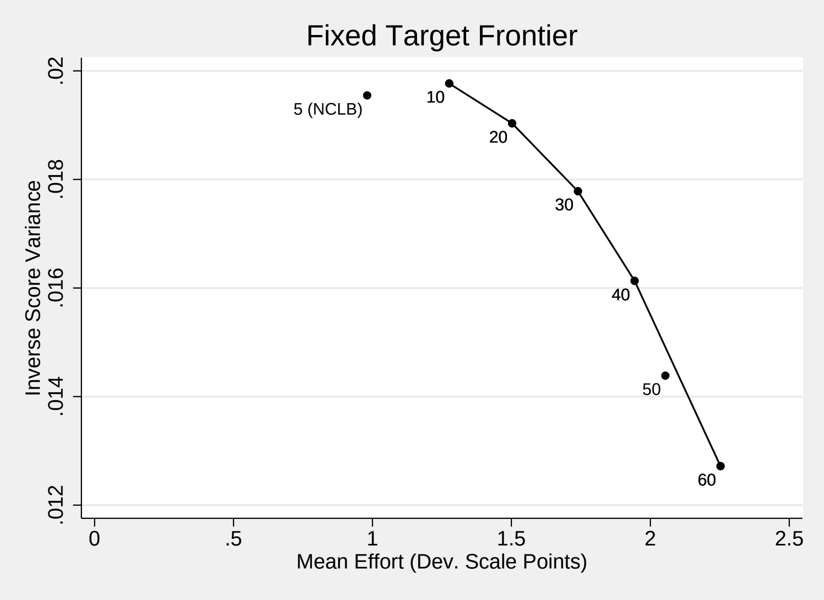
Fixed targets that do not alter the accountability bonus based on student type are the most widespread form of accountability scheme. For such schemes, we show that the choice of the fixed target gives rise to an inherent tradeoff between average teacher effort and test score inequality – a result that is new to the education literature.

To demonstrate this regularity, we first use our counterfactual framework to compute mean effort and a measure of spread – the inverse of the test score variance[56] – for each of a series of fixed targets. Starting at the bottom of the predicted score distribution, we raise the proficiency target up to the real NCLB target and on to higher percentiles in the distribution, plotting the resulting 'mean effort-inverse variance' points to trace out the frontier, as shown in Figure 4. In the figure, we label seven illustrative points associated with seven separate fixed targets, where target labels correspond to target percentile positions in the predicted score distribution.

The frontier shows this clear tradeoff: higher fixed targets lead to greater mean effort but at the cost of higher test score inequality (or lower inverse test score variance). Furthermore, the magnitudes involved are quantitatively significant: moving the proficiency target from the 20th to the 40th percentile of the predicted score distribution increases mean effort by 0.06 standard deviations (in terms of the test score) but at the cost of increasing the test score variance by 18 percent. The figure also indicates that setting progressively higher fixed targets is associated with an increasingly steep tradeoff – for example, raising the target from the 40th to the 60th percentile increases mean effort by only 0.04 standard deviations but raises the test score variance by 27 percent.

To understand why this tradeoff arises, note that when the proficiency target is set relatively

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



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 4 – Fixed Frontier with Homogeneous Bonus Payment

low in predicted score distribution (below the median), increasing it makes a progressively larger mass of students marginal. This creates strong incentives to direct effort to a larger fraction of students, thereby raising mean effort.[57] But higher targets imply that much of the additional effort is directed to progressively better students (those with higher predicted scores), which works to exacerbate performance disparities, raising the test score variance.

Continuing through the distribution, when the target is set relatively high, raising it further makes a progressively *smaller* mass of students marginal and a larger mass of students likely to miss the proficiency target. Without a cost adjustment to keep all schemes comparable, mean effort would decline in these cases because it would be prohibitively costly for teachers to help students meet proficiency standards.[58] In these cases, to equate the cost (proficiency rate) under each scheme with that under the actual NCLB target, the bonus payment needs to be raised so that teachers increase effort and with it, the proficiency rate. Doing so increases mean effort but also increases test score variance, as high-performing students benefit disproportionately from the higher bonus payment due to their marginal position in the incentive strength distribution.

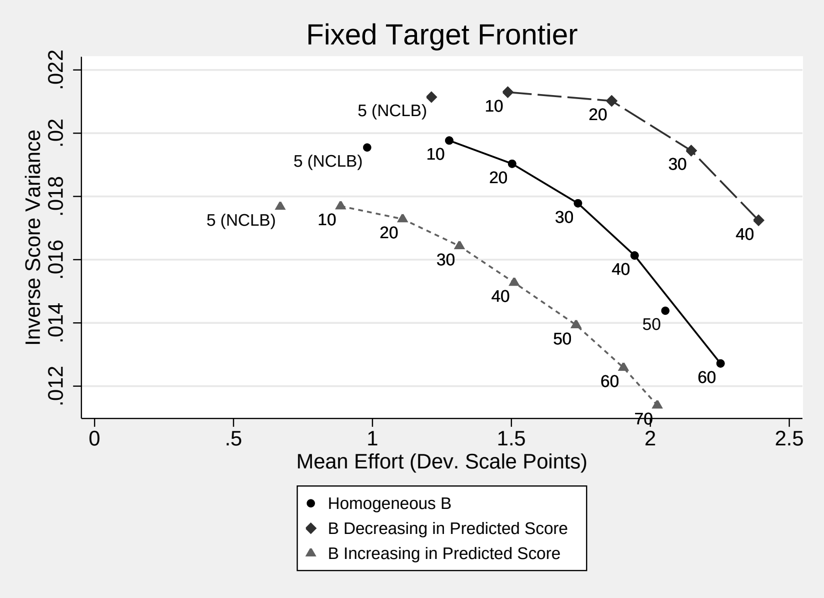
[57] We offer an exact decomposition of the relevant forces at play in Appendix G.1.

[58] That is, teachers would be discouraged by the overly-ambitious standards, responding optimally by exerting less effort (because the probability of target attainment would be very low).

IX.B. Fixed Targets with Heterogeneous Bonus Payments

Next we construct frontiers for the two markedly contrasting heterogeneous bonus-payment regimes described above, favoring low- and high-performing students respectively. Doing so indicates how much outcomes can be altered as a result of redistributing bonus payments.

In particular, Figure 5 shows that the scheme placing more weight on low-performing students dominates the homogeneous bonus payment regime, which in turn dominates the scheme that places more weight on high-performing students. In terms of magnitudes, suppose we hold the proficiency target fixed at the real NCLB value but attach more weight to low-performing students. Doing so increases mean effort by 3.2 percent of a standard deviation and decreases test score variance by 7.5 percent. In contrast, attaching more weight to high-performing students decreases mean effort by 4.3 percent of a standard deviation and increases test score variance by 11 percent.[59]



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 homogeneous bonus payment frontier shown in Figure 4. 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 5 – Fixed Frontiers with Heterogeneous Bonus Payments

To understand why outcomes improve when we place more weight on low-performing students, first consider holding the target fixed at a relatively low level (similar to the real target under

[59] Averaging across all fixed targets, 4.6 percent of a standard deviation more effort and 7.8 percent less variance are achieved under the first regime, and 5.2 percent of a standard deviation less effort and 7.2 percent more variance are achieved under the second regime.

NCLB) and switching to this regime from the 'constant bonus payments' regime.[60] The new regime increases mean effort because it assigns the most weight to low-performing students, essentially 'doubling up' on the already strong incentives for those students. They experience the largest gains in teacher effort as result, which also decreases test score inequality. Continuing through the predicted score distribution, when targets are set relatively high, the new regime creates a tension between the incentive to devote effort to relatively high-performing students (owing to the location of the target in the predicted score distribution) and to low-performing students (due to the greater weight placed on them by the bonus payment system). Without cost equating across regimes, these conflicting incentives result in less overall effort and lower proficiency rates (relative to using the same fixed target but with constant bonus payments). In these cases, the frontier shifts out.[61]

It is clear from Figure 5 that the scheme that assigns more weight to high-performing students is dominated by both other regimes. We discuss the forces that lead to the inward shift of the frontier when switching to this regime in Appendix G.3.[62]

Test Score Gaps Across Demographic Groups: We have shown that the regime offering higher bonus payments for low-performing students dominates the homogeneous bonus payment regime, 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 feasible regime are *substantial.* One way of highlighting the gains is to compute the implied effects of the regime on test score gaps across student subgroups. Table 4 below reports three test score gaps that are of 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, columns (1) and (2) show the observed gap in the data and the gap predicted by our model, respectively.

It is clear our model predicts the *observed* gaps very well. In column (3), we show the percentage of the predicted gap that can be eliminated by switching to the regime where bonus payments are higher for low-performing students. Redistributing bonus payments across students in this way reduces the black-white test score gap by 6.8 percent of its original value, again *without*

[60] Here, we explain the intuition for frontier shifting and provide a decomposition of the relevant forces in Appendix G.2.

[61] This is because we must increase the bonus payment in order to raise effort sufficiently in order to ensure that all schemes are comparable in terms of cost (i.e., the resulting proficiency rate interacted with the bonus payment paid per proficient student).

[62] In this case, when targets are set relatively low in the distribution, a conflict arises between the incentives attached to the location of the proficiency target and those stemming from the nature of the bonus payment, resulting in less overall effort being exerted. When targets are set relatively high in the distribution, teachers face strong incentives to direct effort to high-performing students because of both the location of the proficiency target and the nature of the bonus payment. Average effort increases as a result, but we must scale it back by decreasing the bonus payment to keep all schemes comparable in terms of cost.

changing overall costs. The gap between children of college-educated and less than college-educated parents also falls by a substantial margin – by 5.2 percentage points.

Table 4 – Test Score Gaps and Alternative Schemes

	(1)	(2)	(3)	(4)
			Percent Change in Gap	
Test Score Gap	Observed (SD Units)	Predicted (SD Units)	Fixed Target b decreasing in $\hat{y}$	VA Target with $\alpha = 0.93$
White versus Black	0.78	0.78	-6.8%	+15.3%
College-Educated versus less than College-Educated Parents	0.74	0.75	-5.2%	+12.6%
90th versus 10th Percentile	2.55	2.57	-3.7%	+12.5%

In columns (1) and (2), test score gaps are reported in (student-level) standard deviation units. Column (3) reports the percentage change in the predicted gap (column 2) arising from a switch to the heterogeneous bonus payment regime while continuing to use the real NCLB fixed target of 247 developmental scale points – the fifth percentile of the predicted score distribution. Column (4) reports the percentage change in the predicted gap arising from switching to a VA target regime (with constant bonus payments) using a multiplicative coefficient in the VA target of $\alpha = 0.93$ and an intercept δ that ensure the mean VA target across all students is equal to the fixed NCLB target of 247 developmental scale points.

IX.C. Value-Added Targets

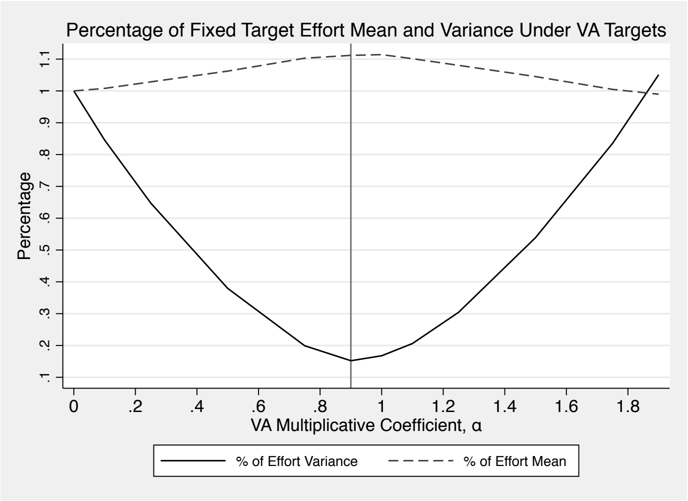
We now explore the properties of value-added (VA) targets. To do so, we exploit the linkage between fixed and value-added targets described in the previous section, whereby each VA target shares the same incentive strength mean (across all students) as a given fixed target but can have a different variance.[63] Our interest centers on how outcomes change as we vary the VA multiplicative coefficient, α, thereby incorporating more student-specific information into the target (through the use of the prior score) and in turn altering the variance of the incentive strength distribution.

The discussion below will reference the incentive strength variance-minimizing value of α, denoted α^*. This variance-minimizing value is straightforward to derive, being given by $cov(\hat{y}_t, y_{t-1})/var(y_{t-1})$.[64] Relative to fixed targets (for which $\alpha = 0$), increasing α up to this critical value α^* *reduces* the variance in incentive strength, causing teachers to apply similar levels of effort to all students; further increasing α past α^* then increases the variance of incentive strength, eventually leading to greater dispersion in effort than under fixed targets.

[63] See Appendix F.3. 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.

[64] Let $var(\hat{y}_t - y_{it}^T)$ denote the variance of incentive strength across all students. For VA targets, we have $y_{it}^T = \delta + \alpha y_{it-1}$, $\forall i$, which allows us to write the variance in incentive strength as $var(\hat{y}_t - y_{it}^T) = var(\hat{y}_t) + \alpha[\alpha var(y_{t-1}) - 2cov(\hat{y}_t, y_{t-1})]$. Taking the partial derivative with respect to α and setting it equal to zero, the variance of incentive strength across all students is minimized at $\alpha^* \equiv cov(\hat{y}_t, y_{t-1})/var(y_{t-1})$. (From another perspective, the critical value α^* is the coefficient from the linear regression of $\hat{y}_t$ on y_{t-1}, which is estimated to be 0.937 in our data.)

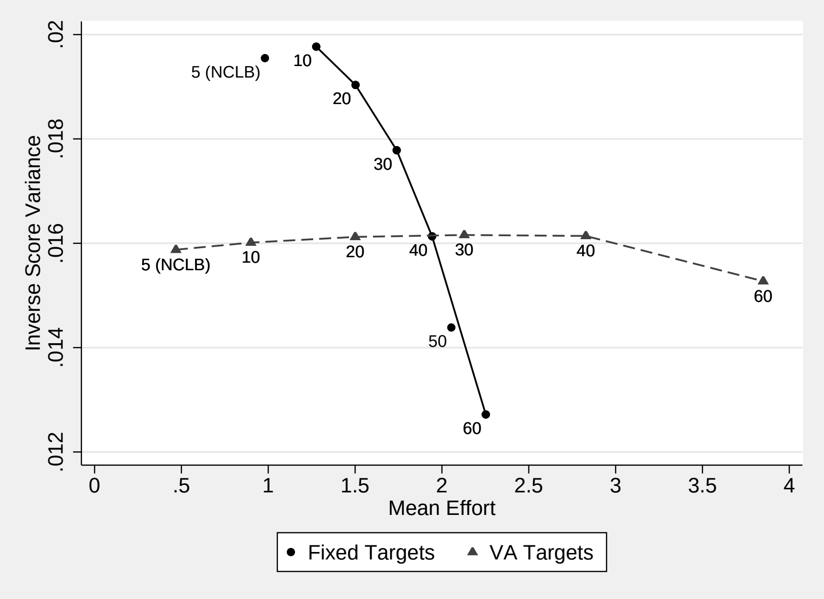
The notion of α^* in place, conditioning proficiency targets on students' prior scores allows policymakers to use VA targets to make a higher fraction of students marginal than under fixed targets, resulting in lower inequality in teacher effort across students. Figure 6 shows that, compared to the fixed target baseline, increasing α toward α^* (the latter indicated by the vertical line) reduces the variance in effort across students progressively, while increasing α above α^* increases the variance progressively (as reflected in the solid line dipping down then rising back up). At the variance-minimizing choice, α^*, VA targets produce less than 20 percent of the effort variance observed under fixed targets. Figure 6 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 fixed targets when α is equal to α^*.



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 vertical line indicates the effort-variance-minimizing α, namely α^*, equal to 0.937 in our data. The dashed line shows 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 shows 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.

Figure 6 – Percentage of Fixed Target Mean and Variance Achieved by VA Targets

Next, Figure 7 shows the mean and variance properties of VA targets in terms of effort (alongside the fixed frontier from Figure 4, for reference), while setting $\alpha = \alpha^*$ for all VA targets. Because there is little variance in effort across students as we change δ and shift the distribution of VA incentive strength, the test score variance is nearly constant across all VA target choices, as all students always receive similar boosts to test scores. The test score variance is higher (the inverse variance is lower) under VA targets than fixed targets when fixed targets are set low in the distribution of student predicted scores (lower than the thirtieth percentile). Here, fixed targets provide stronger incentives to redistribute effort to low-performing students, thus reducing the variance in test scores. The relative inability of VA targets to reduce inequality in these cases is



Notes: The solid line plots the frontier from Figure 4. Each point on the solid line reflects the mean effort and inverse test score variance that prevails under a given fixed target.The point labels correspond to the percentiles the fixed targets in the distribution of student predicted scores. The dashed line plots the frontier arising 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 that the VA intercept δ is chosen to match.

Figure 7 – Fixed and VA Target Frontiers

further documented in Column (4) of Table 4, which shows that switching to VA targets significantly *increases* test score gaps. The white-black test score gap increases by 15.6 percent, for instance, while the gap between students of college-educated and non-college educated parents increases by 12.6 percent. Because VA targets produce relatively little variance in incentive strength across students, they maintain (instead of help to close) performance gaps.

For fixed targets above the thirtieth percentile, the test score variance under VA targets is lower (inverse variance is higher) because the fixed target regimes result in more effort being allocated to relatively high-performing students. VA schemes also result in greater average effort. 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, encouraging teachers to exert more effort. Our simulations indicate that VA targets 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 lower test score inequality.

X. Conclusion

This paper has made three related contributions to the study of incentive design in education.

First, it set out a novel semi-parametric approach for identifying the impact of incentives on effort. This drew on features of the North Carolina context (in particular, the 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 settings to identify teacher effort (detailed administrative data and appropriate policy variation permitting). Doing so is valuable given that effort is typically unobserved and thus difficult to pin down.

The second and third contributions are general in nature. In terms of the second, we proposed a structural procedure based on a flexible model of effort setting, allowing us to identify the primitives underlying the measured teacher effort response. Estimates of the model show that within-classroom tradeoffs in effort across students are important, and that teachers boosted effort following NCLB's introduction.

As our third contribution, the model and estimates formed the basis of a counterfactual framework for measuring the performance of different incentive schemes on a comparable basis. This framework allowed us to assess how effort would change with counterfactual incentives, and to compute the *full distribution* of scores under counterfactual incentive provisions for the first time.

We then 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 emerged from the policyanalysis, each relevant to incentive design in education. First, we showed that fixed targets (of the form taken by NCLB) give rise to a quantitatively significant tradeoff between teacher effort and student test score inequality: higher targets boost average effort at the expense of greater outcome dispersion. Second, the performance of fixed targets can be improved markedly by introducing student-specific *bonuses* that attach higher weight to low-performing students, reducing the black-white test score gap and the score gap between children of college educated versus non-college educated parents 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 90 percent without any sacrifice in aggregate effort.

Stepping back, the counterfactual approach serves as a valuable design tool at a time when states are re-visiting education incentives. By allowing policy makers to gain insight into the distributional consequences of education accountability policies for the first time, it enhances the prospects for using education reforms to combat inequality in a cost-effective manner – an enduring objective of public policy, and one that is especially important today.

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.

Barlevy, Gadi and Derek Neal. 2012. "Pay for Percentile." *American Economic Review*, 102(5): 1805-1831.

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 and Susanna Loeb. 2011. 'School Accountability." *Handbook of Economics of Education*, 3: 383-421.

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.

Loyalka, Prashan, Sean Sylvia, Chengfang Liu, James Chu, and Yaojiang Shi 2018. "Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievements." *Journal of*

Labor Economics, forthcoming.

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

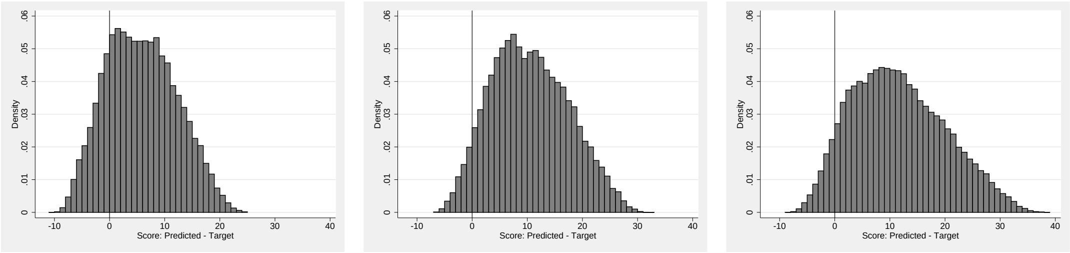
A. Incentive Response: Robustness

This appendix presents evidence supplementing the semi-parametric empirical analysis in Section IV.

A.1. Testing for Bunching

Our semi-parametric approach requires the incentive 'shock' to be exogenous. Indirect light can be shed on whether it is by examining bunching in the distributions of the predicted *ex ante* incentive strength that emerge from applying the proposed recipe, especially in the vicinity of the target.

Figure A.1 plots the grade-specific distributions of our incentive strength measure for 3rd, 4th, and 5th grade mathematics in 2002-03. In each of the panels, the fixed NCLB target occurs at zero, indicated by the vertical line.



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

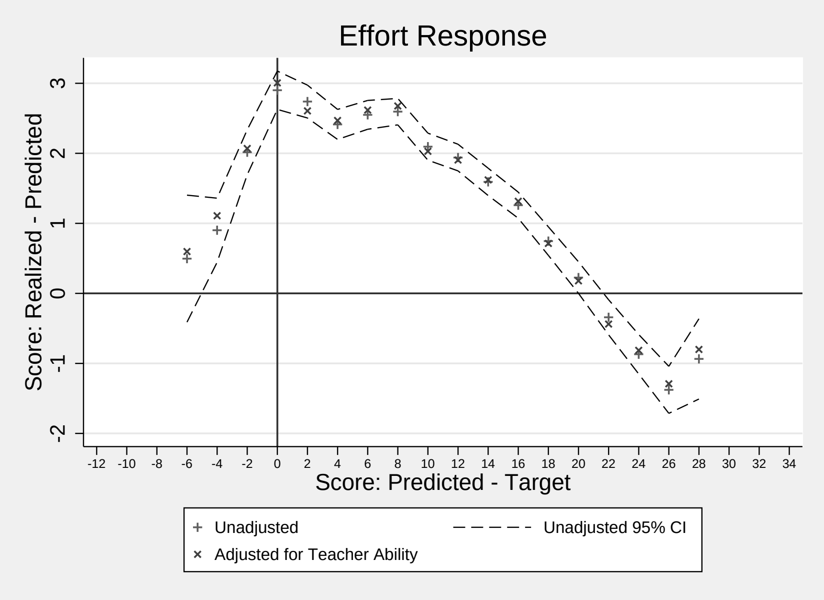
Figure A.1 – Distribution of Predicted Mathematics Scores minus the NCLB Target

Based on the distributions of predicted scores, the figure provides no evidence of bunching – whether around the threshold target or at other parts of the distribution. 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.

A.2. Evidence Supporting the Effort Interpretation

Next, we assess whether the inverted-U pattern in Figure 2(a) might be attributable to educators making adjustments to education inputs other than effort. We consider two potentially important alternative channels that would alter the deployment of school resources. According to the first, schools re-sort students to teachers in response to NCLB's introduction, based on teacher ability – if higher-ability teachers were better able to improve marginal students' test scores, for example. According to the second, schools might sort students to classes differentially based on classroom characteristics, notably class size or classroom homogeneity; they could do so if, for example, schools believed marginal students would perform better in smaller or relatively more homogeneous classes. We take these in turn.

Sorting Based on Teacher Ability. Figure A.2 presents evidence relating to schools sorting students to teachers based on teacher ability in response to NCLB's introduction. The figure is constructed as follows: First, we difference the 2002-03 and 1999-00 profiles of raw means in Figure 2(a) and plot the resulting mean differences along with the associated confidence intervals. We then calculate adjusted means in each incentive strength bin for both 2002-03 and 1999-00 by controlling for the effect of teacher ability.[65]



Notes: This figure presents the profile showing the difference between the 2002-03 and 1999-00 profiles of raw means in Figure 2(a) and the associated confidence intervals *alongside* the difference between the 2002-03 and 1999-00 means adjusted for the effects of teacher ability according to the procedure described in the text.

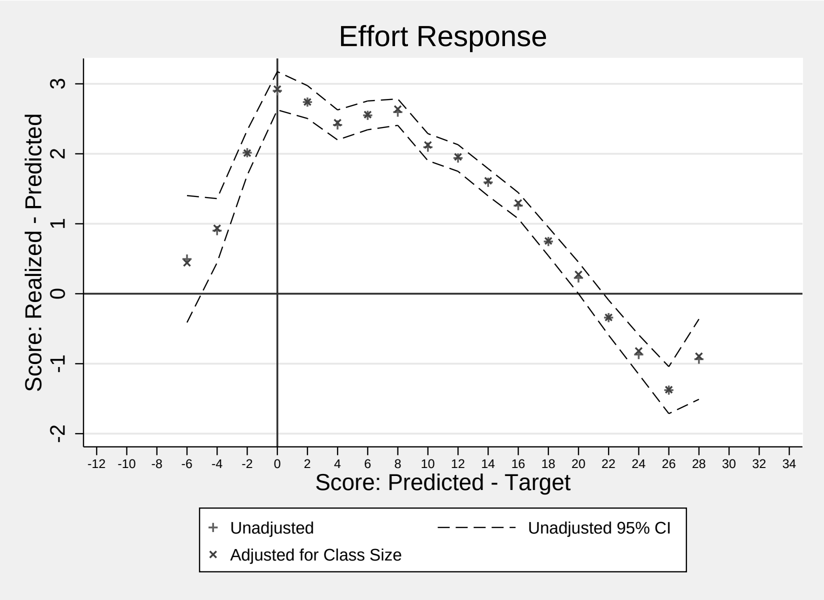
Figure A.2 – The Impact of Differential Sorting by Teacher Ability

The resulting profile of adjusted mean differences falls entirely within the confidence band for the profile of raw mean differences, as shown in the figure. This supports the view that teacher ability does not affect scores *differentially* at any particular point in the distribution, comparing after versus before (2002-03 versus 1999-00). It is therefore unlikely that schools responded to NCLB by sorting students to teachers differentially based on teacher ability in a way that could explain the clear test score patterns we observe.

Sorting Based on Classroom Attributes. Figure A.3 sheds light on whether schools might sort students differentially in response to NCLB's introduction to classrooms based on class size. It does so by repeating the analysis underlying Figure A.2 but adjusting the means in each incentive strength bin by controlling for 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, supporting the view that schools did not respond by sorting

[65] This involves regressing gains above predicted scores on a mutually exclusive and exhaustive set of bin indicators on the horizontal axis, fully interacted with fixed effects for academic years 1999-00 and 2002-03, while also controlling for teacher ability fully interacted with year fixed effects. Our estimates of teacher ability are obtained using the jackknife Empirical Bayes estimator (Chetty *et al.* 2014), which takes data from all years the teacher taught except the year under consideration (2002-03 or 1999-00), thus avoiding the problem of mechanical correlation between test score gains and ability estimates. Within each bin, we then construct the difference in mean gains across years by subtracting the estimated coefficient on the indicator for 1999-00 from the estimated coefficient on the 2002-03 indicator.

students to classrooms differentially based on class size.



Notes: This figure presents the profile showing the difference between the 2002-03 and 1999-00 profiles of raw means in Figure 2(a) and the associated confidence intervals alongside the difference between the 2002-03 and 1999-00 means that are adjusted for the effects of class size.

Figure A.3 – The Impact of Differential Sorting by Class Size

We also examine whether schools responded to NCLB by making classrooms more homogeneous. Creating classes post-NCLB in which students had similar academic preparedness could make it easier for teachers to target instruction at a particular subset of students without necessarily increasing overall effort. To assess this possibility, we examine whether students became more similar within classes by examining the relative changes in within-school and within-classroom variances in prior-year student test scores (as the measure of preparedness). If classrooms became more homogeneous, we would expect the fraction of within-school variation in preparedness that was explained by within-classroom variation to fall in the post-NCLB period.

Figure A.4 plots this fraction – the within-school variance in prior-year test scores 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. Further, there is no discernible change in this fraction in 2002-03, supporting the view that schools did not respond to NCLB by grouping students into more (or less) homogeneous classrooms.

In sum, the observed patterns do not appear to be due to differential sorting on the basis of teacher ability or classroom characteristics. Instead, they are consistent with educators adjusting effort in a targeted manner in response to the introduction of a proficiency-count incentive scheme. This evidence helps justify the model's focus on the effort-setting decision rather than changes in other inputs.

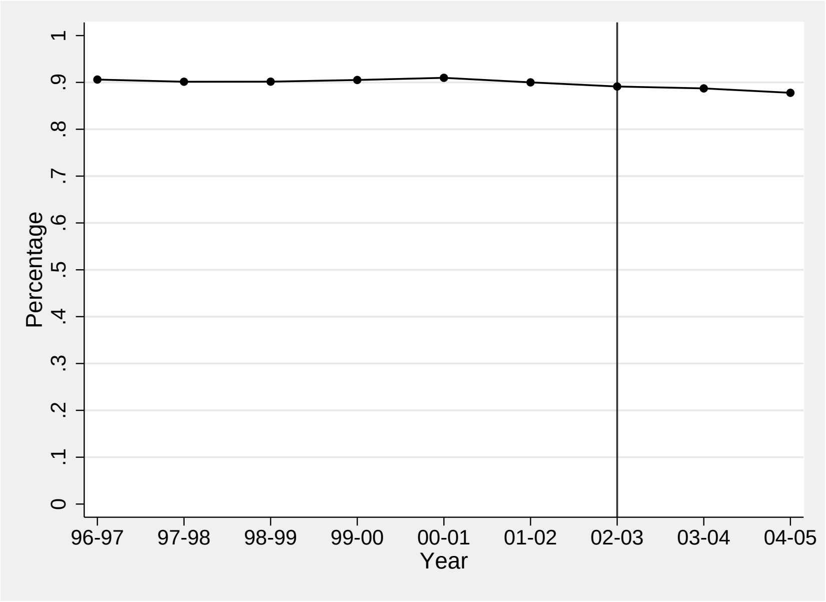


Figure A.4 – Fraction of Within-School Variance in Prior Score Explained by Within-Classroom Variance

Notes: This figure presents, for each academic 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.

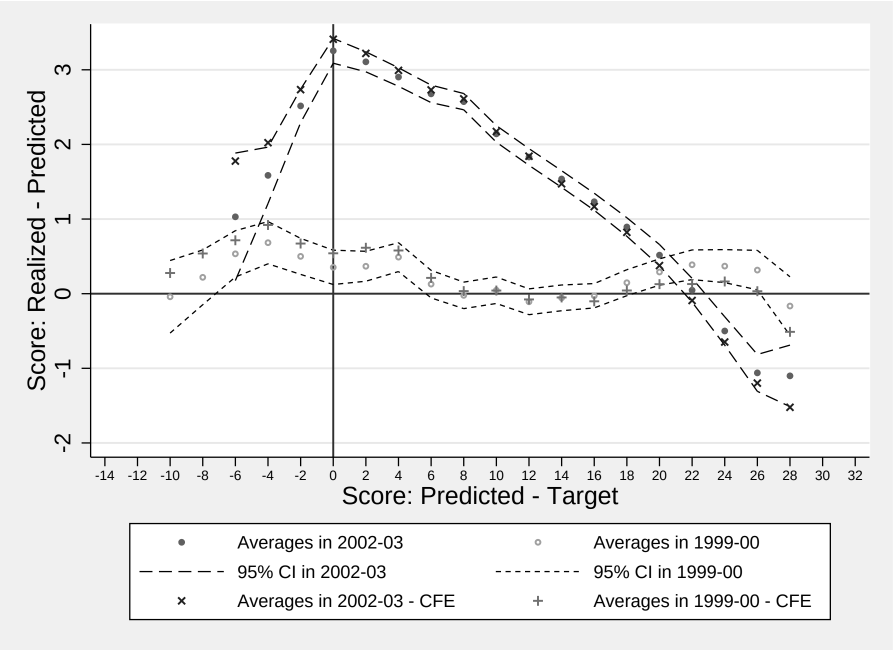
A.3. Evidence Supporting a Teacher Agency Interpretation

The model we develop in Section V treats agency at the *teacher* level, with teachers making decentralized effort decisions based on classroom characteristics. We provide descriptive evidence consistent with this model, indicating that the incentive response identified in Section IV.C occurs almost wholly at the teacher level.

Identifying the Semi-Parametric Patterns Using Within-Classroom Variation: First we show our main results remain essentially unchanged when relying solely on *within-classroom* variation in incentive strength for identification of the semi-parametric patterns presented in Figure 2(a). Figure A.5 plots the raw means in 2002-03 and 1999-00 from Figure 2(a) with the respective confidence intervals, alongside the adjusted means that prevail after removing across-classroom variation.[66] Of note, 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.

Students' Positions in the Classroom Distribution: We also examine whether a student's position in the classroom-level distribution of incentive strength is a better predictor of test score gains than the student's position in the school-level distribution, given the school she attends. The relevant comparison involves students who are in the same position in their school-level incentive strength distribution, but who oc-

[66] The adjusted profiles are obtained by regressing student-level gains above predicted scores on indicators 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.



Notes: This figure presents the raw mean test score gains in 2002-03 and 1999-00 from Figure 2(a), along with the respective confidence intervals. It also plots predicted means in each year identified using only within-classroom variation in test score gains and incentive strength. (See text for a description.)

Figure A.5 – Estimated Effort Using Within-Classroom Variation

cupy different positions in their respective classroom incentive-strength distributions. To test whether such students experienced similar test score gains, we first group students into quartiles of the state-level predicted score distribution and quartiles of their school-level and classroom incentive-strength distributions. 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 low-performing students the most marginal, and Figure 2(a) shows that our estimated effort measure is highest among them. When we regress that effort measure – student i's observed mathematics score (y_i) less her predicted mathematics score ($\hat{y}_i$) – solely on indicators for quartiles of the state-level distribution of $\hat{y}$ in column (1) of Table A.1,[67] we find that effort decreases progressively for students occupying higher positions in the state-level distribution.

Next, we explore whether knowing a student's position in the classroom distribution has any predictive power, over and above the student's position in the state-level distribution. Column (2) of Table A.1 shows that estimated 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, indicating that the position in the classroom distribution is important.[68] In contrast, when testing whether a student's position in the school-level distribution matters, we find the school-level indicators are neither individually nor jointly

[67] Students in the first (lowest) quartile are the omitted baseline category.

[68] To put the magnitudes in perspective, a student who occupies the top quartile in her classroom-level distribution experiences gains that are 0.15 standard deviations lower than those predicted by only her position in the state-level distribution.

Table A.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

Notes: The sample is restricted to students who occupy different quartiles in their classroom- and school-level distributions of incentive strength, respectively (N = 22,199). The dependent variable in each column is our estimated 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.

significant.[69]

In sum, a student's position in his or her classroom incentive strength distribution determines gains over predicted scores in a strong, independent way, while knowing the student's position in her school-level distribution does not have further predictive power (conditional on knowing the student's position in the

[69] See column (3) of Table A.1. 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.

state- and classroom-level distributions). This evidence supports the teacher agency focus in our analysis.

A.4. Response to Target, Not Position in School-Specific Distribution

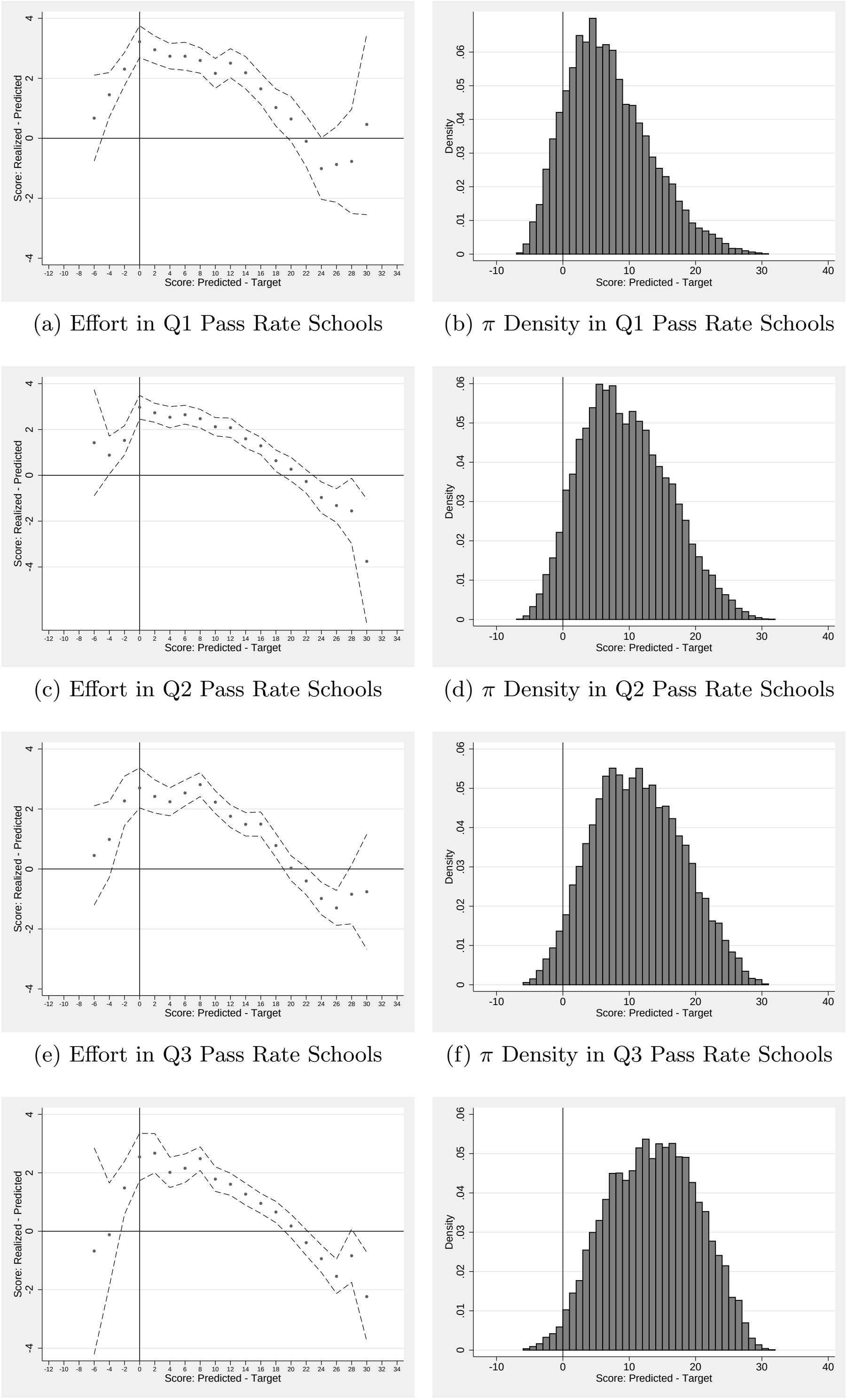
Our maintained hypothesis is that effort is responsive to the incentive strength measure, π, that we have constructed. As an alternative, effort might vary with respect to a student's *relative* position in the predicted score ($\hat{y}$) distribution within his or her school. For example, it is possible that educators responded to NCLB by targeting students at a particular point of the predicted distribution, this point just happening to coincide with the value of $\hat{y}$ where π under NCLB was close to zero.[70] 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 incentive strength (π) equalled zero, then varying π counterfactually to make inferences about competing accountability schemes might would seem unwarranted.

To assess this possibility, we exploit the richness of the administrative data – specifically, by determining the effort responses and corresponding incentive strength densities separately for four types of school, dividing them 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.[71] If schools responded to NCLB by tailoring effort to 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 shifted right across the types of school.

Figure A.6 plots the effort responses and incentive strength 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 a value of π near zero in bottom quartile schools will have a different relative position in the $\hat{y}$ distribution than a student with a value of π near zero in the second, third or top quartile 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 her relative position in the predicted score distribution.

[70] 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 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.

[71] Recall that a student is predicted to pass when $\pi = \hat{y} - y^T > 0$.



(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

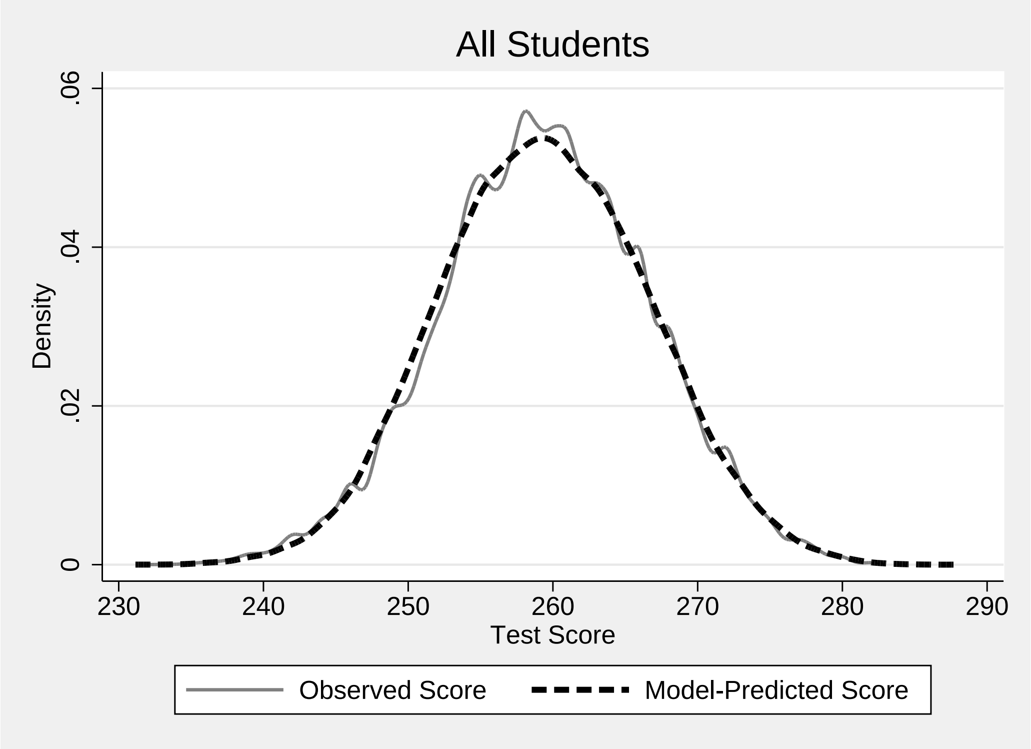
(f) π Density in Q3 Pass Rate Schools

(g) Effort in Q4 Pass Rate Schools

(h) π Density in Q4 Pass Rate Schools

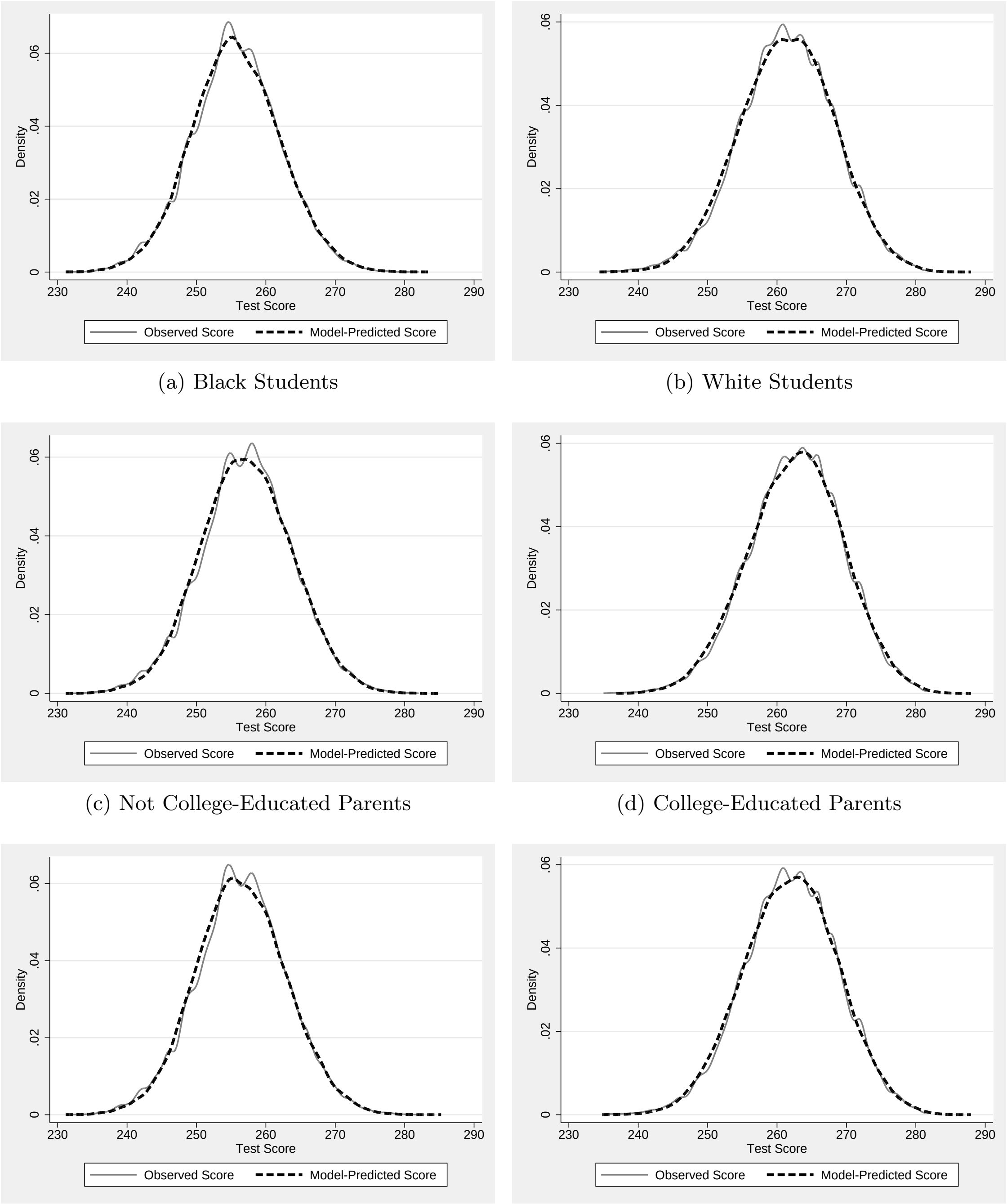
Figure A.6 – Responding to Incentive Strength π rather than the Relative Position of $\hat{y}$

B. Model Fit



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

Figure B.1 – Distributions of Observed and 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 B.2 – Subgroup Distributions of Observed and Model-Predicted Scores

C. Modeling a Response to Multiple Targets

Formal NCLB incentives only apply to the attainment of a single target – the middle proficiency target under North Carolina's pre-existing assessment system – as discussed in Section V. Our model provides additional flexibility, allowing teachers to respond to other pre-existing proficiency targets that may be salient to parents and teachers.

In this appendix, we motivate our multiple-target formulation and provide evidence in support of it, before assessing alternative modeling choices.

As motivation, Table F.1 reports the fraction of students scoring above the low, middle (proficiency), and high (superior performance) targets in each grade and year around the time of NCLB's introduction. The evidence indicates clearly that schools *did* respond to more than the middle target. Specifically, at the top of the distribution, the fraction of students scoring above the *high* target (thereby achieving 'superior performance') increased in 2002-03 by nearly as much or more in each grade than for the proficiency target, even though this was a low-stakes achievement target. In contrast, there is virtually no change in terms of the low achievement target, with nearly 100 percent of students in each grade attaining it throughout; thus, helping students clear this low target was not a relevant margin of adjustment following the introduction of NCLB. Further, in the following year (2003-04), almost no changes in any of the fractions are evident.

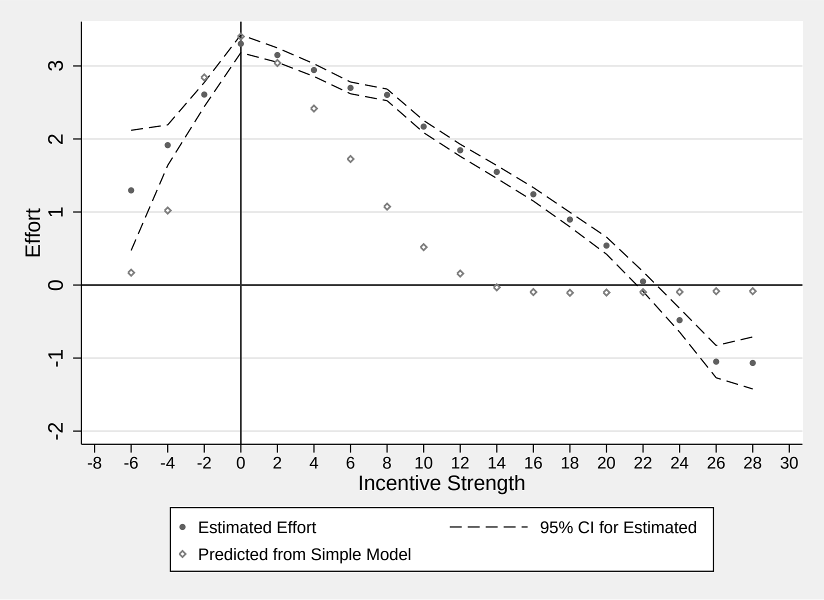
Table C.1 – Fractions of students scoring above low, middle (proficiency), and high (superior performance) targets, by grade and year

	Year		
Grade and Target	2001-02	2002-03	2003-04
Third Grade			
Low	96.9%	98.9%	98.9%
Proficiency	78.3%	89.2%	89.2%
High	37.1%	44.8%	45.1%
Fourth Grade			
Low	99.1%	99.3%	99.2%
Proficiency	89.5%	94.9%	94.6%
High	45.8%	60.5%	60.6%
Fifth Grade			
Low	98.3%	98.9%	99.1%
Proficiency	89.1%	92.7%	93.5%
High	55.6%	63.1%	64.5%

Figure 3 (in Section VII.B) shows what a good fit our multi-target-response model produces with respect to the estimated effort function. We now assess whether a similarly close fit between model predictions and estimates could be achieved through alternative modeling choices. To anticipate, our evaluation of several alternatives considered next will indicate this is not the case, based on the implied fit alongside the estimated effort function.

C.1. A Benchmark: Single-Target Scheme

As a benchmark, consider the simple single-target model, shown in Figure C.1. It is apparent that the fit is poor, particularly in terms of the effort devoted to high-achieving students. A substantially larger σ^2 parameter would broaden out the predicted profile on the right-hand side (via the teacher's first-order condition), as we showed when discussing the comparative static properties of optimal effort. Yet this would be at odds with the value of σ^2 that is identified externally to the teacher's problem, depending as it does only on the difference between estimated and model-implied effort (see Section VI.B).



Notes: This figure presents the estimated effort profile in 2002-03 from Figure 2, along with the 95 percent confidence intervals, and the binned means of model-implied effort for the single-target scheme.

Figure C.1 – Inverted-U Response to NCLB and Model Fit of Single-Target Scheme

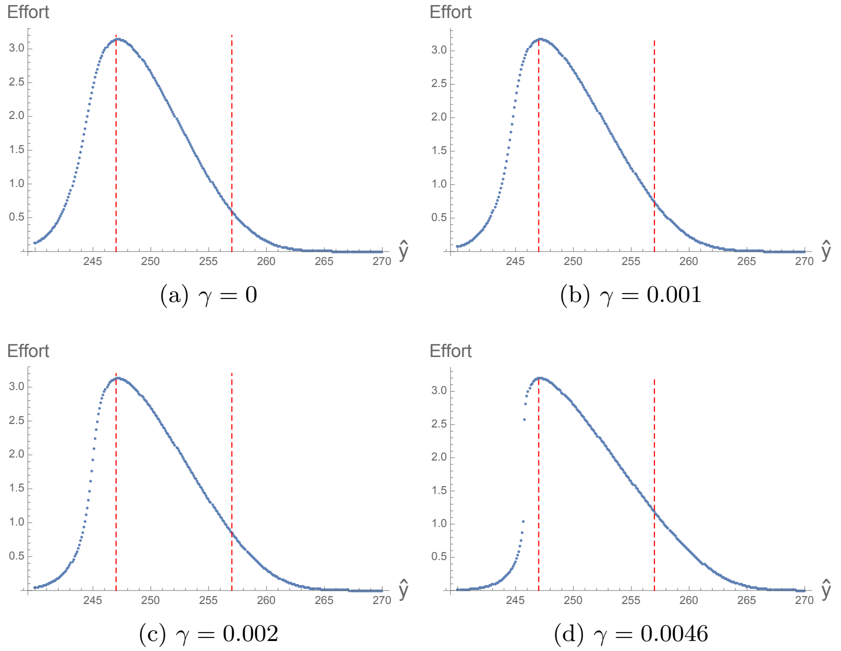
In contrast, as Figure 3 shows, allowing for teachers to respond to the higher target serves to broaden out the model-implied effort profile for high-achieving students, resulting in an exceptionally good match between model and empirics. Next we assess whether alternative modeling choices could achieve the same end.

C.2. Complementarity in Production

One potential way to broaden out the effort profile without changing σ^2 would be to allow for a complementarity in production between student ability and teacher effort, expressed as $y_i = \hat{y}_i + e_i + \gamma\hat{y}_i e_i + \epsilon_i$. If γ were positive, then teacher effort would go further with more able students, resulting in a higher level of optimal effort for those students.

Two considerations count against this alternative formulation. The first relates to the way γ (the coefficient on the interaction term) influences the shape of the implied effort profile. Figure C.2 simulates the effort profile for different values of γ.[72] Relative to the case in which there is no complementarity

[72] To develop intuition, we suppress the cost parameter θ, allowing us to focus on effort setting on a student-by-student basis.



(a) $\gamma = 0$

(b) $\gamma = 0.001$

(c) $\gamma = 0.002$

(d) $\gamma = 0.0046$

Notes: The panels in the figure explore the relationship between optimal effort e^* as a function of $\hat{y}_i$ (the predicted score or student 'ability') and the complementarity parameter γ, capturing an interaction between optimal effort and ability. The dashed line on the left corresponds to the middle (NCLB) proficiency target, while the second dashed line corresponds to a student with a predicted score that is ten points above the middle target (to facilitate comparisons with the estimated effort function in Figure 2).

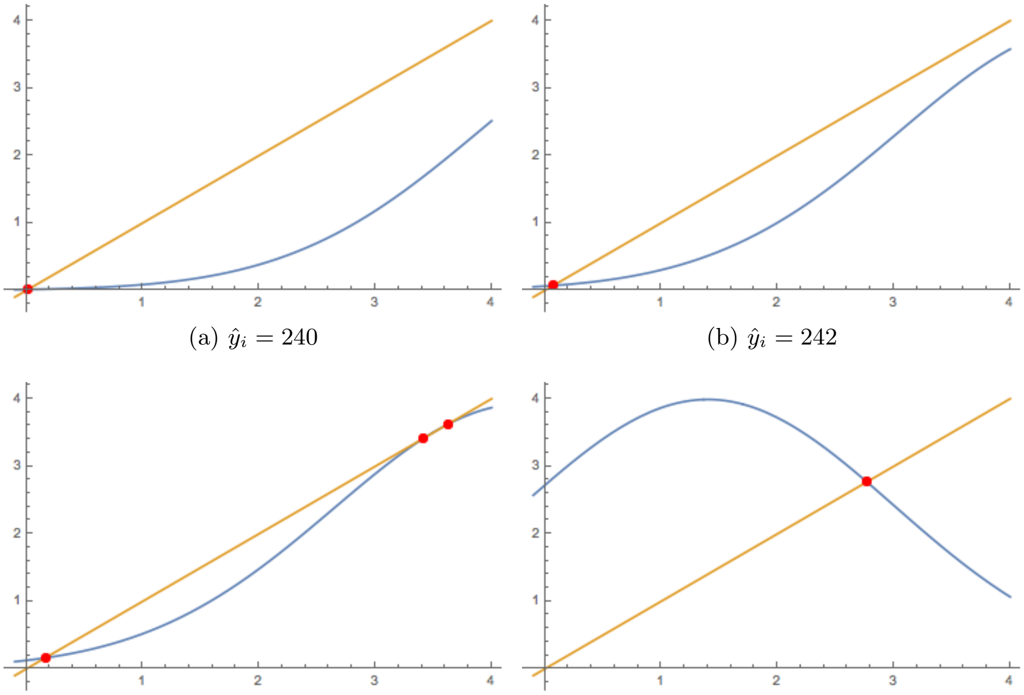
Figure C.2 – Simulation of Effort Function for Different Degrees of Complementarity

(Figure C.2(a), as in our actual model), increasing γ does broaden the right-hand side of the effort profile. Yet broadening the right-hand-side of the model-predicted effort profile to match the estimated effort profile would require a very large value of γ, which in turn would ensure that the two profiles would not match on the left. (Model-predicted effort would decline too quickly, as the bottom two panels show.)[73]

The second consideration relates to the way the marginal benefit and cost curves intersect for different student types (reflected in their predicted scores, $\hat{y}_i$). If the complementarity parameter is strong enough to broaden the right-hand side enough to match the estimated effort profile, as required, this will increase the slope of the marginal benefit curve, at some point making it steeper than the marginal cost curve. If this occurs, it will cause optimal effort to *jump* discontinuously for some value of $\hat{y}_i$ (given the properties of the effort-setting model)[74] – such a jump is illustrated in Figure C.2(d) for $\gamma = 0.0046$; the first value for which the jump is highly pronounced. Yet we see no such evidence of a discontinuity in the estimated effort profile, which rules out complementarities large enough to generate the observed breadth of the effort profile.

[73] The LHS and RHS are relative to the first vertical dashed line in the figure, which represents the NCLB target y^T_M.

[74] Figures C.3(a) and C.3(b) clearly demonstrate this for $\gamma = 0.006$, showing that a low-ability student with $\hat{y}_i = 242$ receives close-to-zero effort, while a near equivalent low-ability student with $\hat{y}_i = 243$ receives approximately 3.5 scale points of effort.



(a) $\hat{y}_i = 240$

(b) $\hat{y}_i = 242$

(c) $\hat{y}_i = 243$

(d) $\hat{y}_i = 250$

Figure C.3 – Simulation of Optimal Effort $(MB = MC)$ for $\gamma = 0.006$

C.3. Peer Spillovers

Another possibility for broadening out the effort profile would be to allow for peer effects in the production technology. Consider, for example, a technology given by $y_i = \hat{y}_i + \rho\bar{\hat{y}} + e_i + \epsilon_i$, where the test score of student i also depends on the average ability of the other students in the classroom $(\bar{\hat{y}})$. If such peer effects are positive as is plausible (i.e., $\rho > 0$), the model cannot generate a broader effort profile on the right-hand side of the horizontal (incentive strength) axis, where needed. To understand why, note that the peer effects term effectively lowers the proficiency target faced by each student in a classroom-specific way, where y^T becomes $y^T - \rho\bar{\hat{y}}$. Reducing the effective target creates incentives to direct more effort toward students with low values of $\hat{y}$, but we require stronger incentives for students with high values of $\hat{y}$ to generate a broader effort profile on the right-hand side. Thus, introducing peer effects into the model does not allow us to replicate the observed patterns in the data.

C.4. Prediction Error

Suppose that teachers face uncertainty about student ability, observing a noisy signal of $\hat{y}_i$, given by $\tilde{y}_i = \hat{y}_i + \nu_i$, where (for illustration) ν is independent of ϵ and is normally distributed with mean μ_ν and variance σ_ν^2. From the teacher's perspective, the test score technology would be given by

$$y_i = \tilde{y} + e_i + \epsilon_i = \hat{y}_i + e_i + \underbrace{\nu_i + \epsilon_i}_{\eta_i} = \hat{y}_i + e_i + \eta_i. \tag{13}$$

Because ϵ and ν are both normally distributed, their sum, $\eta = \epsilon + \nu$, is also normally distributed, and the relevant mean and variance parameters in the teacher's problem are given by $\mu_\eta = \mu_\epsilon + \mu_\nu$ and $\sigma_\eta^2 = \sigma_\epsilon^2 + \sigma_\nu^2$, respectively.

The variance σ_ϵ^2 is pinned down by the observed difference between estimated and model-implied effort, as noted above, and it is too small to generate the required broadening of the effort profile on the right-hand side. While the variance term associated with prediction error (σ_ν^2) could close the gap between the two, its value would need to be exceedingly large to do so: the total variance of the teacher-observed signal ($\tilde{y}$) would need to be more than three times larger than the observed variance of $\hat{y}$, and the interquartile range of $\tilde{y}$ would be nearly two times wider than that of $\hat{y}$. Such values are implausible. In addition, it is likely that teachers are in a position to predict the performance of fourth grade students quite accurately, leading us to rule out prediction error as a viable alternative explanation.

D. Properties of the Optimal Effort Solution

In this appendix, we further analyze the properties of optimal effort, building on the discussion in the main text. There, we noted that effort in our model does not have a closed-form solution (see Section V.B). Further, the model does not generically have a unique solution. Yet we are able to compute a global maximum numerically, necessary for the estimation and simulation exercises we carry out in the main text. We are also able to show (via simulation) how the global optimum is likely to translate into a unique set of corresponding model parameters, as we now discuss.[75] We will focus on the subset of parameters $\tilde{\beta} \equiv (\frac{b_M}{\psi}, \frac{b_H}{\psi}, d_M, d_H) \subset \beta$. This is because σ^2 is identified externally from the teacher's problem, and θ is identified from the two values of $\hat{y}$ for which the estimated effort profile turns negative (conditional on average classroom effort and σ^2) – see Section VI.B.

A simulation approach can be used to show that a given effort profile corresponds to a unique set of parameter values $\beta \equiv (\tilde{\beta}; \theta, \sigma^2)$. Here, we appeal to the estimated effort profile from Section IV.C, combined with the model's first-order conditions. Specifically, for each predicted score $\hat{y}_i$, we replace e_i^* in the expression for the first-order condition with the corresponding value from the estimated profile.

Formally, let $e^*(\tilde{\beta}; \pi_i)$ be the optimal effort exerted by a teacher of student i, with $\pi_i \equiv \hat{y}_i - y_M^T$. Recalling equation (7) and using the fact that the M and H targets differ by eleven developmental scale points ($y_H^T = y_M^T + 11$), optimal effort satisfies

$$\frac{b_M}{\psi} f(d_M - \pi_i - e_i^*) + \frac{b_H}{\psi} f(d_H - \pi_i - e_i^* + 11) - e_i^* - \theta \sum_{j=1}^{N_c} e_j^* = 0\,,$$

where $\theta = 0.0075$ and $\sigma^2 = 15.702$ from Section VII, $\sum_{j=1}^{N_c} e_j^*$ is known from applying the effort profile to the average teacher, and e_i^* is used as shorthand for $e^*(\tilde{\beta}; \pi_i)$.

The optimizing solutions under consideration are restricted to the subset of parameter vectors that satisfy the first-order condition, given that the empirical effort profile is taken as the truth. There are many such vectors, but this subset is far smaller than all possible parameter vectors, most of which are unable to recover the effort pattern we observe. Limiting ourselves to the smaller feasible subset of parameter vectors makes the problem tractable. Uniqueness is then defined in the following way: there does not exist $\tilde{\beta}' \neq \tilde{\beta}$ such that $e^*(\tilde{\beta}; \pi_i) = e^*(\tilde{\beta}', \pi_i)\ \forall\ i$ for a given incentive strength. In words, two different feasible parameter vectors cannot both yield the same global maximum in terms of the teacher objective across all students types.

We use a four-dimensional grid search (using different initial guesses for each of the four parameters) to solve for candidate parameter vectors that satisfy the function. While there are approximately forty discrete student types contained within the support of our estimated effort function (given that developmental scale points are integers); for tractability, we select ten representative points from that function to match in the simulation that follows. Doing so takes advantage of information from the estimated profile about how effort

[75] Fixed point theorems can be used to establish uniqueness analytically for lower dimensionality problems (e.g., if there were only one target to consider, with two parameters). They do not apply in our setting, however, given that it involves four interdependent parameters, two of which (d_M and d_H) shift the effective distance between the two known targets.

declines away from each of the peaks. In particular, any interplay between the response to each target will be stronger for values of $\hat{y}$ in between the two targets than values to the left of the M target or the right of the H target. By considering points in each region of the predicted distribution, our simulation exercise is able to account for such interplay.

We discretize the parameter space in terms of what to consider for initial guesses. In particular, we consider eleven initial points for each of the four parameters. The set of initial value guesses is represented by the $\varphi(\cdot)$ function, where $\varphi(\frac{b_M}{\psi}) = \varphi(\frac{b_H}{\psi}) = \{0, 10, \ldots, 90, 100\}$ and $\varphi(d_M) = \varphi(d_H) = \{-10, -8, \ldots, 8, 10\}$. The parameter space of $\tilde{\beta} \equiv (\frac{b_M}{\psi}, \frac{b_H}{\psi}, d_M, d_H)$ is then $\varphi(\frac{b_M}{\psi}) \times \varphi(\frac{b_H}{\psi}) \times \varphi(d_M) \times \varphi(d_H)$. Given the number of data points we match (i.e., the number of non-linear equations – ten in our case), initial value vectors are constructed from the resulting grid values and then used as a starting point when solving the system of non-linear equations with the gradient method. The grid is reasonably comprehensive, with 14,641 ($= 11^4$) initial starting points for the vector $\tilde{\beta}$.

Across all 14,641 initial value combinations, there are only two candidate vectors that maximize the teacher objective globally: $\tilde{\beta}_1 = (30.49, 23.91, 3.13, 1.74)$ and $\tilde{\beta}_2 = (23.91, 30.49, 12.74, -7.87)$. Both solutions result in a teacher objective value of 335.886 and an error rate of 0.5686.[76] The first solution ($\tilde{\beta}_1$) is reached from 11,281 (or 77%) of the initial starting points, while the second solution ($\tilde{\beta}_2$) is reached from 2,071 (or 16%) of them. Around 7% of starting points are unstable, converging to a different objective value each time.[77] In short, there is no convergence of the gradient method for these initial grid vectors.

At first glance, it might seem problematic that there are two viable candidates, $\tilde{\beta}_1$ and $\tilde{\beta}_2$, which both imply the same objective value. However, closer inspection reveals that they are simply mirror images of each other, a possibility which our simulation does not rule out. In particular, $[\frac{b_M}{\psi}]_1 = [\frac{b_H}{\psi}]_2 = 30.49$ and $[\frac{b_H}{\psi}]_1 = [\frac{b_M}{\psi}]_2 = 23.91$, while $[d_M]_2 = 12.74 = [d_H]_1 + 11$ and $[d_H]_2 = -7.87 = [d_M]_1 - 11$. That is, the second candidate simply swaps the M and H labels, producing exactly the same solution. Thus, fully 93% of initial starting points converge to the same global solution of $\tilde{\beta}_1 = (30.49, 23.91, 3.13, 1.74)$.

We note that while $\tilde{\beta}_1$ is qualitatively similar to our estimated parameters in Section VII, it is not an exact match.[78] This is not surprising, as our simulation routine abstracts from how students are distributed across classrooms and exploits only a subset of the information contained in the estimated effort function. Nevertheless, this simulation exercise shares enough in common with the full estimation routine for the uniqueness argument to carry over.

[76] The error rate is $\sqrt{g_1^2 + \ldots + g_{10}^2}$, where $g_i \equiv \frac{b_M}{\psi} f(d_M - \pi_i - e_i^*) + \frac{b_H}{\psi} f(d_H - \pi_i - e_i^* + 11) - e_i^* - \theta \sum_{j=1}^{N_c} e_j^*$ is evaluated using the candidate parameter vector and effort is taken from point i of the empirical effort function. It measures how closely the parameters satisfy the first-order conditions, across all ten points under consideration from the effort function, with an error rate of zero implying that they are exactly satisfied.

[77] They also have associated error rates that are at least five times larger than the two stable candidates, which indicates that the resulting ‘solutions’ do not actually satisfy the first-order conditions.

[78] Recall that the estimates are $(\frac{b_M}{\psi}, \frac{b_H}{\psi}, d_M, d_H) = (36.3, 24, 3.19, 1.63)$.

E. Computation Appendix

In this appendix, we describe how effort is computed in the model presented in Section V.

For a given value of the parameter vector $\beta \equiv (\frac{\mathbf{b}}{\psi}, \mathbf{d}, \theta, \sigma^2)$, we solve for the optimal level of effort devoted to each student (denoted by $e^*\left(\beta; \hat{y}_i, \mathbf{y^T}, \hat{\mathbf{y}}_c\right)$) by maximizing the corresponding teacher's objective function in equation (6) with respect to the full vector of optimal effort levels for all students in the class $\{e^*\left(\beta; \hat{y}_1, \mathbf{y^T}, \hat{\mathbf{y}}_c\right), \ldots, e^*\left(\beta; \hat{y}_{N_c}, \mathbf{y^T}, \hat{\mathbf{y}}_c\right)\}$. As shown in Section V.B, this results in N_c first-order conditions for each classroom, given by equation (7), where the unknown variables are the N_c optimal effort values; the first-order conditions are interdependent within classrooms, as the effort devoted to any given student in the class depends on the effort received by all other students. The first-order conditions are independent across classrooms, however, implying that solving for the full distribution of optimal effort amounts to solving N_c first-order conditions simultaneously in each classroom.

In practice, we carry out this exercise in Matlab by maximizing the teacher's objective function in each classroom with respect to the N_c effort levels. We do so using Matlab's built-in unconstrained minimization package *fminunc*, while supplying both the gradient vector and Hessian matrix to ensure that the solution vector to the first-order conditions in equation (7) indeed maximizes the teacher's objective.[79] We loop over all classrooms in the data, maximizing a new teacher's objective function on each iteration, until we recover the full distribution of optimal effort levels across all students.

[79] Because *fminunc* is a minimization routine, we apply it to the negative of the teacher's objective function, ensuring the recovered solution maximizes the objective function.

F. Simulation Appendix

This appendix provides background to the counterfactual simulations, relating to the setting of targets and bonus payments, and the cost-equating procedures we use.

F.1. Counterfactual Fixed Targets

We construct counterfactual fixed targets designed to cover the full predicted score distribution. The targets are measured on the End-of-Grade Mathematics test developmental scale, following the protocol under NCLB. In total, we consider 17 fixed targets, starting from 237 developmental scale points and increasing the target by an increment of 2 points on each iteration. The set of fixed targets is thus

$$y^T \in Y^f = \{237, 239, 241, \ldots 247, \ldots 263, 265, 269\}.$$

Table F.1 shows the mapping between each developmental scale point target and the corresponding percentile in the predicted score distribution (of $\hat{y}$); the actual NCLB test score proficiency target (in bold in the table) is set at 247 developmental scale points, corresponding to the fifth percentile of the predicted score distribution. The table makes clear this set of fixed targets covers the entirety of the predicted score distribution, aside from the very top.

Table F.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

F.2. Heterogeneous Bonus Payments

We consider two different regimes in which bonus payments are heterogeneous across students: In the first case, the student-specific bonus payment is given by $b^L(\hat{y}_i) = b_M \frac{(\hat{y}_{\max}+1-\hat{y}_i)}{\hat{y}_{\max}-\hat{y}_{\text{med}}+1}$, where $\hat{y}_{\max}$ is the maximum value of $\hat{y}_i$ across all students in the state and $\hat{y}_{\text{med}}$ 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_M is the per-student bonus payment from before, which is now scaled by the student-specific weight $w_i = w^L(\hat{y}_i) = \frac{(\hat{y}_{\max}+1-\hat{y}_i)}{\hat{y}_{\max}-\hat{y}_{\text{med}}+1}$. In second case, the student-specific bonus payment is given by $b^H(\hat{y}_i) = b_M \frac{(\hat{y}_i - y_{\min}+1)}{\hat{y}_{\text{med}}-\hat{y}_{\min}+1}$, 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_M is scaled by the student-specific weight $w_i = w^H(\hat{y}_i) = \frac{(\hat{y}_i - y_{\min}+1)}{\hat{y}_{\text{med}}-\hat{y}_{\min}+1}$.

To illustrate the form of these two heterogeneous bonus payment parameterizations, Figure F.1 below shows each of them as a function of predicted scores along with the baseline homogeneous bonus payment case in which $w_i = 1, \; \forall \; i$, along with the density of the predicted score distribution in the background.

The two chosen parameterizations of the heterogeneous bonus payment are convenient for three reasons. First, they cover two informative extremes: in the first case, the bonus is highest for the worst students and lowest for the best students, while in the second case, the opposite is true. Second, they ensure that the original payment b_M is being multiplied by a number that ensures cost control: in the first case, b_M 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_M multiplied by 1.) Third, since the parameterizations are determined in a data-driven way, they can be calculated in any dataset.

F.3. Counterfactual Value-Added Targets

In general, value-added ('VA') targets are set based on information contained in students' prior scores. As such, there are many potential ways of constructing them.[80] To keep the analysis tractable, we restrict attention to counterfactual VA targets that use students' prior scores from only one subject (mathematics) and that are linear in those scores. Thus we write a student i's specific VA target at time t as $y^T_{it} = \delta + \alpha y_{it-1}$, where y_{it-1} is i's mathematics score at $t-1$.

As noted in the main text, fixed targets can be viewed as special cases of VA targets – specifically, where $\delta = y^T$ and $\alpha = 0$. By setting the δ parameter appropriately, we can ensure that a given fixed target has a VA counterpart that delivers the same mean for the distribution of incentive strength $(\hat{y}_{it} - y^T_{it})$ across all students as the fixed target does. The VA counterparts will generally have smaller variances because the use of the prior score allows student-specific targets to be set that can make many more students marginal. By considering several different multiplicative coefficients (α) for each fixed target, and adjusting the intercept (δ) to match the mean of the fixed target, we are able to explore the effects on student outcomes of both mean shifts of, and variance changes to, the incentive strength distribution.

In our simulations, along those lines, we take each fixed target in the set Y^f in turn. By varying $\alpha \in \Omega = \{0, 0.1, 0.25, 0.5, 0.75, 0.9, 1, 1.1, 1.25, 1.5, 1.75, 1.9\}$, we allow the prior score to play a progressively more important role. For each α, we then 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^f. (For

[80] For example, during the 1990s, North Carolina's own ABCs program used *both* prior mathematics and reading scores linearly when setting targets for either subject, while South Carolina's accountability program used both scores and incorporated linear, quadratic, and interacted terms.

example, suppose we are matching real NCLB fixed target, 247. In that case, we set $\delta = 247 - \alpha\bar{y}_{t-1}$, which implies that the mean VA target is also 247.)

We conduct this exercise for each fixed target in Y^f, looping through $y^T \in Y^f$. For each y^T, we then loop through $\alpha \in \Omega$. In doing so, for a given $\alpha \in \Omega$, we pick $\delta(\alpha) = y^T - \alpha\bar{y}_{t-1}$, thus ensuring that the mean VA target is equal to y^T (the mean fixed target).[81]

F.4. Cost-Equating Procedure

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

$$Q_R = \frac{b_M \sum_{i=1}^{N_t} 1\left(\hat{y}_{it} + e^*\left(w_i, y^R_{it}; \widehat{\beta}, \hat{y}_{it}, \hat{\mathbf{y}}_c\right) + \epsilon_{it} - y^R_{it} \geq 0\right)}{N_t}, \tag{14}$$

where N_t is the total number of students in the state, $\widehat{\beta} \equiv [\widehat{\frac{b_M}{\psi}}, \widehat{d}_M, \widehat{\theta}, \widehat{\sigma}^2]$,[82] 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, $y^R_{ir} = y^R$, $\forall\ i$, 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 $y^R_{it} = \delta^R + \alpha^R y_{it-1}$.

To explain the cost-equating procedure, we first focus on the case where bonus payments are constant across students and $w_i = 1$ for all students. While the parameter b_M is not separately identified from m in our model, we can without loss of generality normalize b_M to one and interpret the estimated ratio $\widehat{\frac{b_M}{\psi}}$ accordingly. To equate costs across target regimes, we define $b(k) = kb_M = k$ to be the original bonus payment value multiplied by a constant $k > 0$. Multiplying b_M by k implies evaluating the effort function in equation (14) at the argument $k \cdot \widehat{\frac{b_M}{\psi}}$ instead of $\widehat{\frac{b_M}{\psi}}$ and multiplying the sum in the numerator by k (instead of b_M, 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 $y^T = 247$.[83]

With this notation in place, we use the following procedure to equate the cost that prevails under the set of targets R to the value 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 in a direct way the amount paid per student who passes. This is seen by recognizing that the sum of the indicator variables in equation

[81] To see this, note that we have $y^T_{it} = \delta + \alpha y_{it-1} = y^T - \alpha\bar{y}_{t-1} + \alpha y_{it-1}$, thus implying that the mean value of y^T_{it} (across all students) is y^T. Because both the fixed and VA targets have the same mean, it then necessarily follows that the mean of incentive strength under the fixed target is equivalent to the mean of incentive strength under the VA target. Letting $\bar{\hat{y}}_t$ denote the mean predicted score across all students in time t, mean incentive strength under both the fixed and VA targets is given by $\bar{\hat{y}}_t - y^T$.

[82] When bonus payments are homogeneous, we set $w_i = 1$ for all students.

[83] 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.

(14) is multiplied by a different value each time k adjusts. The second effect comes from the impact of changing k (equivalently, the bonus payment) on teacher effort decisions, which is made clear by the effort function in equation (14) being evaluated at the argument $k \cdot \widehat{\frac{b_M}{\psi}}$ instead of $\widehat{\frac{b_M}{\psi}}$. Increasing k increases costs by raising both the payment per each passing student and incentivizing 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).

Heterogeneous Bonus Payments: Modifying the Cost-Equating Procedure

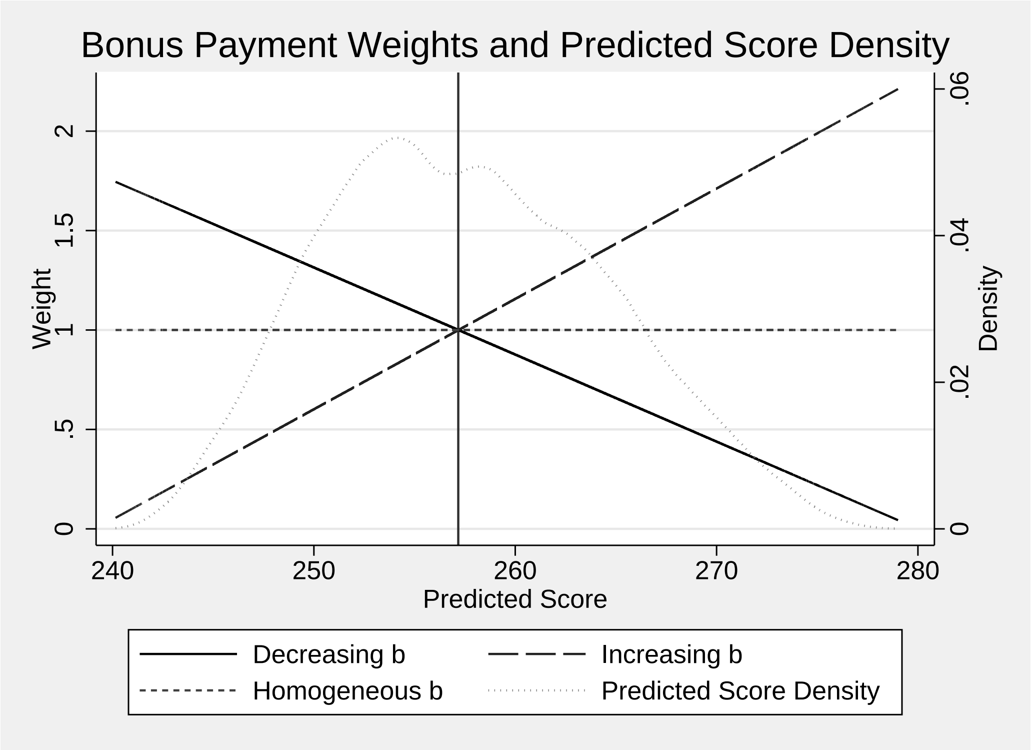
Under heterogeneous bonus payment regimes, average costs under the scheme that places more weight on low-performing students and the scheme that places more weight on high-performing students are determined by

$$Q_R^L = \frac{b_M \sum_{i=1}^{N_t} w^L(\hat{y}_{it}) 1\left(\hat{y}_{it} + e^*(w^L(\hat{y}_{it}), y_{it}^R; \widehat{\beta}, \hat{y}_{it}, \hat{\mathbf{y}}_c) + \epsilon_{it} - y_{it}^R \geq 0\right)}{N_t} \tag{15}$$

and

$$Q_R^H = \frac{b_M \sum_{i=1}^{N_t} w^H(\hat{y}_{it}) 1\left(\hat{y}_{it} + e^*(w^H(\hat{y}_{it}), y_{it}^R; \widehat{\beta}, \hat{y}_{it}, \hat{\mathbf{y}}_c) + \epsilon_{it} - y_{it}^R \geq 0\right)}{N_t}, \tag{16}$$

respectively. For each heterogeneous bonus payment case, we cost-equate across regimes to Q^* using the same methodology described above: we normalize b_M to 1, multiply it by k, and adjust k until costs equate to Q^*.



Notes: This figure shows the three bonus payment (b) regimes as functions of predicted scores, with the density of the predicted score across all students in the background. 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 F.1 – BONUS PAYMENT WEIGHTS AND PREDICTED SCORE DENSITY

G. Fixed Target Frontier Appendix

In this appendix, we provide decompositions of the factors that explain movements along, and shifts of, the fixed target frontier under constant and heterogeneous bonus payments, respectively.

G.1. Constant Bonus Payments

In Section IX, we showed that increasing the fixed target to higher points in the distribution of the predicted score causes aggregate effort to increase but at the expense of also increasing test score inequality (variance). In this appendix, we provide a more detailed discussion of the movements along the fixed target frontier under constant bonus payments, showing that the effort-variance tradeoff reflects the operation of two effects. The first, which we label the 'distributional effect' (or 'DE') for convenience, captures the response of teacher effort as the target is set higher in the predicted score distribution. The second – labelled the 'cost-equating effect' (or 'CEE') – reflects how outcomes change when we adjust the bonus payment to equate costs across all target regimes. For each target we consider, the magnitude of each effect is calculated relative to the baseline mean effort and inverse test score variance that prevail under the real NCLB target.

Which force dominates in determining the shape of the frontier depends in an intuitive way 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 reflects the DE. Increasing the fixed target while it still *below* the median makes a progressively larger mass of students marginal, creating sharper incentives for more students[84] and leading to higher mean effort. But increasing the target also makes progressively *better* students marginal, implying that low-performing students receive relatively little effort, exacerbating performance inequality and increasing test score variance. Setting progressively higher targets in this range therefore increases mean effort and raises test score variance (thus lowering the inverse variance), resulting in the downward-sloping frontier shape depicted in Figure 4.

When the target is above the median of 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.[85] The DE leads to reductions in mean effort because the targets make it prohibitively costly for teachers to help their students meet proficiency standards. In such cases, the bonus payment b_M must be increased to raise effort and equate costs with the benchmark regime given by the real NCLB target. Doing so increases both mean effort and the test score variance (so decreasing the inverse variance), as high-performing students benefit disproportionately from the higher bonus payment, owing to their marginal position in the incentive strength distribution.

The relevant forces at play are further illustrated in Table G.2, which shows the precise magnitudes of

[84] This is true empirically. Suppose we define a student as 'marginal' if her predicted score is within 4 developmental scale points of the target – a relatively tight window. Then the fractions of marginal students when targets are set at the 5th, 20th, and 40th percentiles of the predicted score distribution are 0.19, 0.34, and 0.4, respectively. (Other candidate 'marginal' windows lead to similar patterns.)

[85] 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 75th and 95th percentiles of the predicted score distribution are 0.31 and 0.15, respectively, while the fractions of non-marginal students who are predicted to fail are 0.50 and 0.84.

the DE and CEE (on both mean effort and inverse variance) for several representative fixed targets.

Table G.2 – 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)	DE	CEE	DE	CEE
10	0.23	0.07	-0.00006	0.0003
20	0.40	0.12	-0.00087	0.0003
30	0.49	0.27	-0.00223	-0.0005
40	0.50	0.47	-0.00369	-0.0003
60	0.32	0.96	-0.00593	-0.0009

In columns (1) and (2), we present the impacts of the Distributional Effect (DE) and Cost-Equating Effect (CEE), as defined in the text, on mean effort, respectively, as we move along the frontier in Figure 4 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, though reporting effects of the DE and CEE on the inverse of the test score variance.

G.2. Heterogeneous Bonus Payments Decreasing in Students' Predicted Scores

Next, we explain why outcomes improve when we switch to a regime that places more weight on low-performing students by distinguishing between two effects that cause the frontier to shift out: the 'bonus payment effect' (or 'BPE'), constituting the change that results from switching the bonus payment structure but without ensuring cost equivalence; and the CEE, representing the subsequent change that results from equating costs.[86] For each target we consider, the magnitude of each is calculated relative to the prevailing mean effort and inverse test score variance under that same target in the homogenous bonus payments case.

Holding the target fixed at the real NCLB target and switching regimes from homogeneous to heterogeneous bonus payments increases mean effort because the new regime assigns the most weight to low-performing students, essentially 'doubling up' on already strong incentives for those students.[87] In addition, because students at the bottom receive 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 outward shift in the frontier – that is, to cause both an increase mean effort and a reduction in test score variance.

For higher targets on the frontier, it is the CEE that increases mean effort when switching bonus payment regimes. At relatively high proficiency targets,[88] the BPE creates tension between the incentive to devote effort to (relatively) high-performing students and the incentive to devote effort to low-performing

[86] There is no DE here because the target does not change. The description we offer explains *shifts* in 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.

[87] For example, a the real NCLB target, the mean effort gain (from switching bonus payment regimes) among students below the median of the predicted score distribution is 0.49 developmental scale points (7 percent of a standard deviation). In contrast, students above the median lose 0.03 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).

[88] More specifically, the BPE results in mean effort increases for targets up to the 20th percentile of the predicted score distribution, after which point the CEE is needed to increase mean effort and generate the frontier's outward shift.

students due to the heterogeneous bonus payments. As a result of the BPE, high-performing students are allocated less effort than under the homogeneous bonus payment regime, whereas low-performing students receive more, the net effect being an overall reduction in average effort.[89] 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. In these cases, the CEE more than offsets the BPE, increasing mean effort above the original value under heterogeneous bonus payments and resulting in an outward shift in the frontier.

Table G.3 quantifies the forces that cause the shift out of the froniter in Figure 5. For each target reported on the frontier, we present the precise magnitudes of the BPE and the CEE when switching bonus payment regimes.

Table G.3 – Decomposition of the Bonus Payment and Cost-Equating Effects when 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	0.23	0.00	0.0015	0.0000
10	0.21	0.00	0.0015	0.0000
20	0.16	0.20	0.0012	0.0008
30	-0.04	0.45	0.0005	0.0011
40	-0.35	0.79	0.0003	0.0008

In columns (1) and (2), we present the BPE and the CEE on mean effort, respectively, that occur in Figure 5 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 effects of the DE and CEE on the inverse of the test score variance.

G.3. Heterogeneous Bonus Payments Increasing in Students' Predicted Scores

As Figure 5 makes very clear, the scheme that assigns more weight to high-performing students is dominated by both other regimes. We now discuss this regime in more detail; for brevity, we do not provide a full decomposition of the effects in terms 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 a summary of the resulting adjustment.

As is the case for the regimes described above, the mechanics of the adjustment depend on the location of the proficiency target. 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). At high proficiency targets,

[89] For example, when the target is set at the 30th percentile of the predicted score distribution, students with predicted scores above the median receive 0.10 developmental scale points less effort (on average) as a result of the BPE, while students with predicted scores below the median receive 0.19 developmental scale points more effort. These effects are equivalent to 0.013 and 0.026 standard deviations of the test score, respectively.

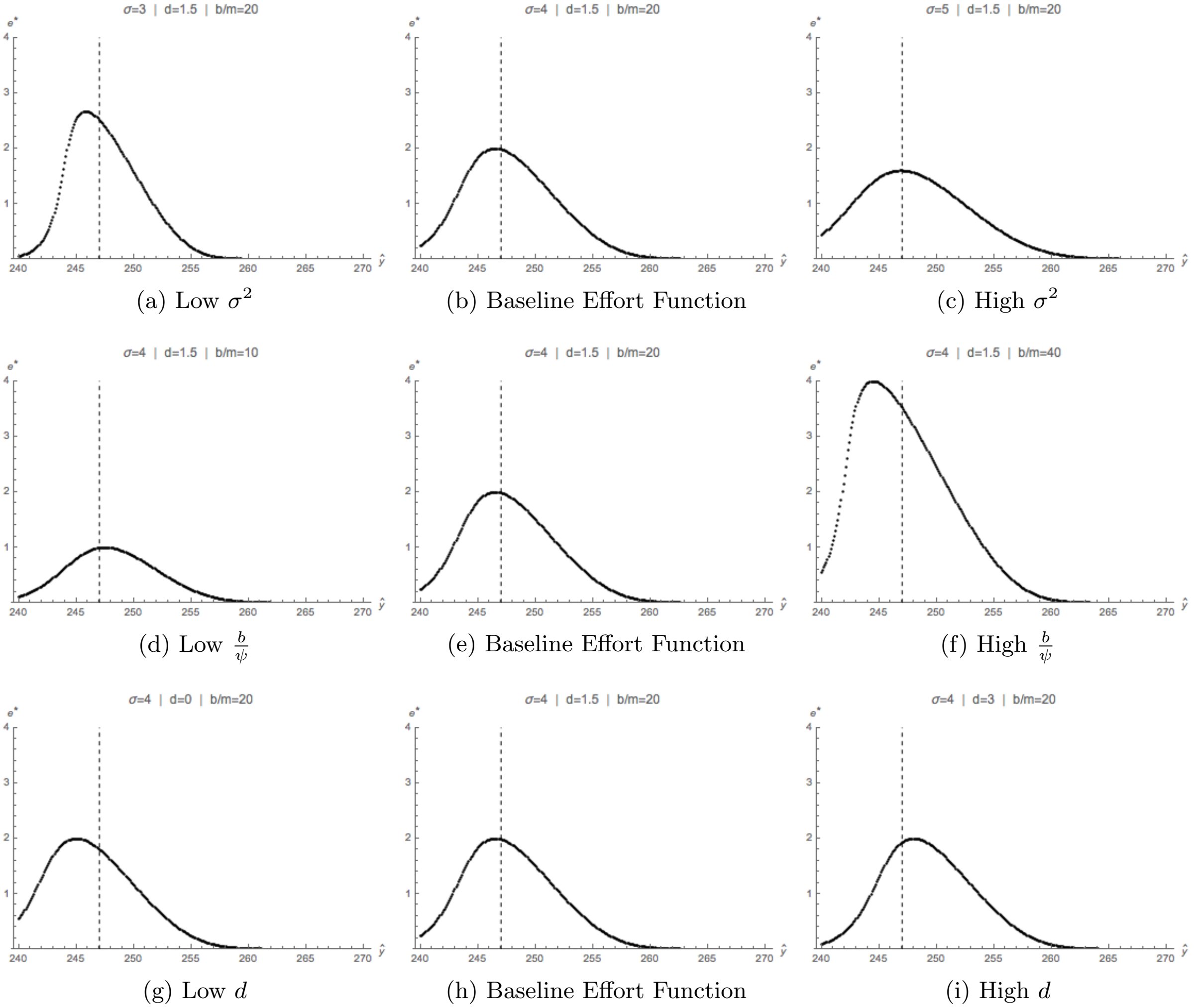
both proficiency target incentives and bonus payment incentives are strongest for students who are high in the predicted test score distribution. These students receive the largest amount of extra effort, while low-performing students experience the largest reduction. Because the strongest students experience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

H. Supplemental Tables and Figures

Table H.1 – Student-Level Descriptive Statistics

	Structural Analysis Sample
Mathematics Score	259.51
	(7.17)
Reading Score	152.89
	(8.05)
College-Educated Parents	0.26
	(0.44)
Male	0.50
	(0.50)
Minority	0.40
	(0.49)
Disabled	0.05
	(0.22)
Limited English Proficient	0.03
	(0.16)
Free or Reduced-Price Lunch	0.45
	(0.50)
Sample Size	89, 271

Notes: Summary statistics are calculated over all fourth grade student observations from 2002-03.



(a) Low σ^2

(b) Baseline Effort Function

(c) High σ^2

(d) Low $\frac{b}{\psi}$

(e) Baseline Effort Function

(f) High $\frac{b}{\psi}$

(g) Low d

(h) Baseline Effort Function

(i) High d

Figure H.1 – Comparative Statics of Optimal Effort With Respect to Changing the Model Parameters

Notes: The panels in the figure illustrate the response of the optimal effort profile to changes in model parameters. Panels (a) through (c) show how the *spread* of the profile increases as σ^2 rises, panels (d) through (f) show how the *horizontal location* of the profile's maximum shifts rightward as d rises, and panels (g) through (i) show how the *height* of the profile increases as $\frac{b}{\psi}$ rises.