A Quantitative Framework for Analyzing the Distributional Effects of Incentive Schemes*

by Hugh Macartney, Robert McMillan, and Uros Petronijevic†

May 2021

Abstract

This paper develops the first quantitative framework for analyzing distributional effects of incentive schemes in public education. The analysis is built around a hump-shaped effort function, estimated semi-parametrically using exogenous incentive variation and rich administrative data. We identify key primitives that rationalize this effort function by estimating a flexible teacher effort-choice model. Both the model and parameter estimates are necessary components in our counterfactual framework for tracing the effects of alternative accountability systems on the entire test score distribution, with effort adjusting endogenously. We find widespread schemes that set a fixed target for all students give rise to a steep performance-inequality tradeoff. Further, counterfactual incentive policies can outperform existing schemes for the same cost – reducing the black-white test score gap by 7% (via student-specific bonuses), and lowering test-score inequality across students by 90% (via student-specific targets). Our quantitative approach opens up new possibilities for incentive design in practice.

Keywords: Incentives, Effort, Accountability Scheme, Education Production, Test Score Distribution, Inequality, Conditional Average Treatment Effect, Semi-Parametric, Counterfactual, Education Reform

JEL Classifications: D82, I21, J33, M52

*We would like to thank Joe Altonji, Peter Arcidiacono, David Deming, Giacomo De Giorgi, David Figlio, Chris Flinn, Caroline Hoxby, Lisa Kahn, Kory Kroft, Lance Lochner, Derek Neal, Rich Romano, Eduardo Souza-Rodrigues, Aloysius Siow, Chris Taber, and seminar participants at Duke University, UCSD, the University of Florida, the NBER, NYU, SITE, Western, Wisconsin, and Yale for helpful comments and suggestions. Marc-Antoine Chatelain, Elaine Guo, Guan Yi Lin, and Hammad Shaikh 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 – see Figlio and Loeb (2011) for an informative survey. 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 to address distributional concerns. The latter are especially relevant as differences in outcomes while in school are known to perpetuate, fuelling lifelong inequality. Yet from a policy perspective, no existing education research allows policy designers to assess the effects of different possible accountability reforms on the entire student performance distribution in a systematic way.

In this paper, we propose a quantitative framework to fill the gap. Our framework is built around an effort function that maps formal incentives under the control of the policy maker into teacher effort – a key input to education production. Teacher effort is typically unobserved, yet it should (as a productive input) be reflected in observed output, and be responsive in predictable ways to incentives. When accountability incentives change, we show how effort can be backed out from observed changes in output without making further assumptions about the effort-setting process, yielding the desired effort function.

We uncover the shape of the effort function in practice by leveraging rich administrative data from North Carolina following all public school students over time, along with exogenous incentive changes arising from the introduction of No Child Left Behind ('NCLB'), a federal accountability system. NCLB created well-documented incentives to focus on students predicted to score close to a fixed passing threshold rather than students further away, in line with proficiency count schemes elsewhere (in Texas, for example). We capture these non-uniform incentives with a continuous measure of incentive strength for all students, equal to their predicted distance to the passing threshold, with incentives being stronger closer by. Then we estimate the profile of conditional average treatment effects of the policy throughout the incentive strength distribution, conditioning on the incentive measure. The resulting profile has a pronounced hump, peaking where incentives should be most intense and declining on either side. Further, supporting evidence indicates that this pattern is not due to other relevant inputs being altered (including class sizes or teacher and student classroom assignments), consistent with teachers adjusting their effort.[2]

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

[2] 'Effort' in our analysis will be taken to refer to unobserved discretionary actions, distinct from other education

We show this effort function can be rationalized using a flexible teacher effort-choice model, adapted to features of the public school system in North Carolina; the further structure will be necessary for our counterfactual approach. The model makes explicit how teacher effort is influenced by incentives – formal and possibly informal – on the benefit side of the choice problem, while the convex cost of teacher effort is flexible, allowing for cost spillovers across students. The model's benefit and cost parameters can be credibly identified, as we demonstrate, and the resulting model estimates fit the data closely, including features of the data not targeted in the estimation routine. These estimates indicate that teachers responded to the introduction of NCLB proficiency by boosting effort on a targeted basis, taking account of the NCLB incentive threshold as well as an informal low-stakes performance target established by the state prior to NCLB. Further, we find the marginal cost of effort is increasing in the effort devoted to other students in the classroom.

At the heart of the analysis is a simulation framework for conducting policy-oriented counterfactuals, drawing on both the model and the parameter 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 output we produce. Our emphasis will be on feasible schemes that are related to existing accountability systems, the most widely-used being schemes (such as NCLB) that set fixed performance targets for all students in a given grade; we also consider value-added schemes whose targets can be tailored to depend on prior scores of individual students.

Using our counterfactual policy framework, we first compare the relative performance of existing schemes throughout the distribution, and then compute the effects of feasible schemes yet to be implemented. 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. We focus on two summary measures under each reform: the average effort received by students[3] (as an indicator of efficiency) and the reciprocal of the variance in test scores (as an indicator of inequality). The resulting counterfactual output then provides a menu of options enabling policymakers to select their favored scheme according to their preferences, as defined over efficiency and inequality.[4]

Three new findings emerge from the counterfactual analysis. First, widely-used fixed target regimes (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 a broad range of possible targets in turn, showing that as the fixed target moves up the predicted test score distribution, average effort increases at the expense

inputs, which alter test scores in ways attributable to incentive variation (see Section IV).

[3]This is equivalent to the average test score in our setup, given a monotonic relationship between effort and scores.

[4]This positive emphasis contrasts with the normative approach in the optimal contracting literature; see Mirrlees (1975) and related theoretical studies. Implementing the optimal contract would raise feasibility issues beyond the scope of the current analysis.

of creating a wider outcome distribution. This makes target setting consequential, depending on policy makers' preferences over average performance (effort) versus inequality. Here, our framework offers 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. These gains all come at no extra expense, as we apply a cost-equating procedure to ensure that all schemes under consideration cost the same amount.

Third, student-specific *targets* allow policymakers to reduce inequality without sacrificing average effort. We show that switching from fixed to student-specific value-added targets (where the target is a function of the student's prior performance) reduces inequality 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 enough 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 40th 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 also improving average student performance without increasing costs. Further, by allowing policy makers to gain insight into the distributional consequences of alternative education accountability 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 study to existing research. Section III describes the incentive variation and the administrative data used in the analysis. Section IV presents our method for uncovering the effort-incentive strength relationship, along with estimates and supporting evidence. Section V develops a model of effort setting that can rationalize the estimated function, with Section VI describing 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, primarily in the personnel, labor and education fields. First is 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 incentives.[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 clear 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 a new 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 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 especially concerning 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 approach we propose for recovering the effort function uses an incentive strength measure building on prior work, including Reback (2008), Neal and Schanzenbach (2010), and Deming *et*

[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.

al. (2016). Two somewhat subtle aspects of our chosen measure are worth noting. First, because it is continuous, it 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 baseline effort predictions that plausibly exclude the treatment effect of the reform (see 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); in turn, distinctive patterns of output are related to prevailing incentives to infer how unobserved effort must be set, and the model is taken to the data to estimate benefit and cost parameters governing worker decisions – the 2009 study by Copeland and Monnet provides an excellent example.[10] 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 impacts of alternative accountability incentives (and the targets and rewards/penalties they entail) on the resulting distribution of educational outcomes. Such design issues come naturally to mind when considering the various education accountability schemes that have been used, including proficiency schemes (such as NCLB) and value-added schemes whose targets condition on prior student scores.[11]

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, our counterfactual policy framework provides a feasible means of constructing alternative incentive schemes that can be fed into the policy analysis, and it allows their impacts to be measured on a comparable basis by equating costs. Further, their effects on the *full distribution* of relevant outcomes can also be traced for the first time, allowing

[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); those studies do not have access to a pre-reform period.

[9] For an illuminating survey of the personnel literature more generally, see Prendergast (1999).

[10] They 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. Structural methods have been used in recent papers in labor economics – for instance, to shed new light on the effects of imperfect competition. See studies by Lamadon, Mogstad and Setzler (2021) and Kroft, Luo, Mogstad and Setzler (2021).

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

convenient summary measures to be computed,[12] according to the preference of the analyst or policy maker.

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 (based on proportions attaining the targets). We focus on the AYP targets shared by all students in a given grade, treating those as a reasonable approximation to the prevailing incentives under NCLB.[13] Doing so provides a potent source of across-student variation in teachers' incentives to devote extra effort, as we will demonstrate.[14]

The federal NCLB program was introduced on top of the state's pre-existing school-based 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 pre-existing incentives under the ABCs contrast sharply with those under NCLB: the former give incentives to exert reasonably uniform effort throughout the distribution, while the latter are distinctly non-uniform. Prior research does not provide guidance in terms of how to treat possible interactions between the two – a practical issue we address in the estimation below.

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 individual score distribution,[15] students achieving proficiency status when their test

[12] An experimental study by Loyalka *et al.* (2019) 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).

[13] The same stance is taken in Neal and Schanzenbach (2010).

[14] 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 studies (Reback 2008; Deming *et al.* 2016) have used student subgroup membership to identify accountability pressure across students and schools.

[15] 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.

scores met or exceeded the second target. Upon NCLB's introduction in 2002-03, North Carolina used the second target as the NCLB proficiency standard, although the first and third may have had some continued salience for educators and parents, a possibility we will examine.

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; this feature 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 in the main dataset, where classes are self-contained and the most accurate matching of students to teachers is possible.[16] 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, 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 mathematics test score scale. These test scores 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), 6 percent are learning-disabled, only 3 percent are limited English-proficient, and 44 percent are eligible for free or reduced-price lunch. Around a quarter of students have college-educated parents.

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

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.

IV. Recovering the Effort Function

In this section, we first estimate the conditional average treatment effects of NCLB, giving rise to a hump-shaped profile with a distinctive right-skewed shape. We then present evidence that supports a 'teachers boosting effort' interpretation of the policy response reflected in this profile, drawing on our administrative data. This evidence will motivate the teacher effort choice model we develop in the next section, which serves to rationalize the policy effects and also provides the foundation for our counterfactual analysis.

IV.A. Policy Responses to Accountability

Measured responses to heightened accountability have already been documented widely in the literature, as noted in Section II. In our setting, we measure the test score responses to the introduction of NCLB in 2002-03 throughout the distribution of incentives. To do so, we adapt standard methods and leverage our rich administrative data.

To set out the approach, we define some terms at the outset. The accountability threshold is denoted by the target score, y^T. The actual score for a given student i is y_i, to be contrasted with the student's *ex ante* predicted score, denoted $\hat{y}_i$. The NCLB test score *response* at the individual

level can then be defined by the difference between the actual score for student i and the student's predicted score (or $y_i - \hat{y}_i$ in the above notation).

The predicted score plays an important role in the analysis. The specific notion we have in mind is the test score that the student would have earned in the absence of any policy, in the year NCLB was introduced. To that end, we devise a prediction algorithm using data and test score determination based on the years prior to NCLB being introduced. The predicted score is obtained by applying the following 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.[17]

2. Then predict student performance in 2002-03 using the saved coefficients from the first step, along with student covariates in 2002-03 and prior test scores from 2001-02, giving $\hat{y}_i$. (This predicted score will, by construction, exclude any response to NCLB and is *ex ante* in that sense.)

This approach provides accurate test score predictions. As an indication, we group students into narrow bins based on their predicted scores, then show that the bin means of the actual scores of students coincide almost exactly with the corresponding predicted scores defining each bin. This is true throughout the predicted score distribution – see Figure A.1. Some prediction error still remains (as reflected in the error bands in the figure), reflecting the uncertainty in the test score process: our investigations show that an R-squared of around 0.73 for the prediction algorithm is about as high as it is possible to go.

We capture the relevant incentives by defining a natural measure of *incentive strength*. This is student-specific, given by the difference between the student's predicted score, $\hat{y}_i$, and the accountability threshold, y^T. For convenience, we label the incentive strength measure $\pi_i \equiv \hat{y}_i - y^T$, noting that this 'predicted minus target score' difference is continuous and also straightforward to construct for each student i in our administrative dataset.

To understand the incentives associated with this measure, threshold schemes of the form taken by NCLB should (as noted) lead educators to focus on students predicted to be close to the proficiency threshold. Thus incentives should be strongest where students are most marginal ($\pi \approx 0$), becoming weaker as the absolute value of the distance from the target increases, in either direction. The corresponding test score effects should in turn be non-uniform, peaking where students are most marginal.

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

To assess whether this pattern emerges in our data, we will estimate the conditional average treatment effects of NCLB (in terms of test scores) for different values of the incentive strength measure. The test score response for student i can be expressed using standard 'potential outcomes' notation as $y_i(1) - y_i(0)$, where $y_i(1)$ denotes the actual score and $y_i(0)$ is the counterfactual score in the absence of the reform. Here we will assume that the counterfactual score can be represented by the test score prediction from above, $\hat{y}_i$, plus a prediction error, ϵ_i. The rationale for using this prediction is that it excludes – by construction – any impact of NCLB by using estimated relationships from before the introduction of the policy, combined with covariates in the year the policy was introduced. It should provide our best possible estimate of each student's score in the (counterfactual) absence of the policy.

Define the effect of the policy at the individual level as $e_i \equiv y_i(1) - y_i(0)$. Thus, substituting for the counterfactual and noting that $y_i(1) = y_i$, the test score response relates to the policy effect for student i as follows:

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

We will allow the policy effect to be heterogeneous in terms of our incentive measure π, writing $e_i = e_i(\pi_i)$. Further, we assume that the treatment effect is homogeneous when conditioning on a given value of π, so that $e_i(\pi_i) = e(\pi_i = \pi)$, common to all students with the same incentive strength. Thus, under this minimal structure, we have

$$y_i - \hat{y}_i = e(\pi_i = \pi) + \epsilon_i \,, \tag{1}$$

where ϵ_i is also implicitly linked, via the characteristics of student i, to a corresponding value of π. The estimand of interest, the *conditional average treatment effect* of the policy, is then written $E[y_i - \hat{y}_i | \pi_i = \pi]$ for any given value of incentive strength π.

In principle, taking the conditional expectation using (1) for any π will yield the conditional average treatment effect $e(\pi)$, assuming $E[\epsilon|\pi] = 0$. While this assumption is not directly testable, we will offer grounds below for believing that it is reasonable in our application.

Our interest will be in recovering the entire profile of conditional average treatment effects of the policy with our incentive measure (π) on the horizontal axis. To estimate the profile in practice, we average the individual test score responses across all students within small intervals (bins) of the incentive measure, defined by the value of π at the bin's midpoint.[18] Applying our binning procedure, the estimated gains across the distribution are plotted in Figure 1 (with a bin size of 2),

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

using fourth grade test scores in 2002-03, post-NCLB.[19] The test score response shows a distinctive inverted-U shape, peaking where incentives under NCLB should be strongest (at zero) and declining on either side of that. It also has an asymmetric shape, which we will seek to rationalize based on the model in the next section.

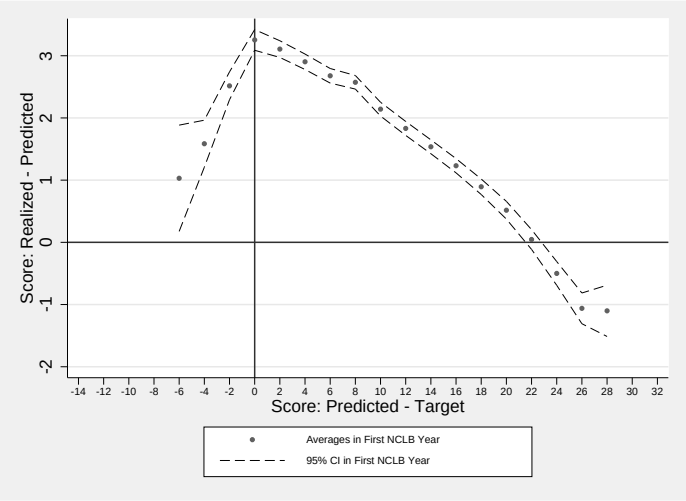


FIGURE 1 – EFFORT RESPONSES

Our identifying assumption for the conditional average treatment effect profile is that no other policy or input accounts for the estimated pattern. This amounts to assuming that the error term has mean zero for each value of π (as stated), and is uncorrelated with the main policy effect, given by $e(\pi)$. These assumptions are plausible in this context, for two reasons: first, we will show that no other policies would have the distinctive hump-shaped profile we have recovered from the treatment effects estimation exercise, peaking where incentives should be strongest and declining on either side.[20] Second, we will find no evidence of shifts in other inputs that would generate anything like the distinctive profile.

A seemingly appealing placebo comes to mind, essentially by computing the policy response for a year *prior* to NCLB's introduction. We note, however, that the resulting profile would not provide a direct (or conclusive) representation of the relevant counterfactual. Any deviations from a flat profile for a preceding year do not imply that the correct counterfactual placebo in the post-reform year would not itself have a flat profile: the profile will capture unmeasured factors operating in the pre-NCLB year at different points in the distribution. At the same time, any such factors influencing scores are accounted for in our post-reform profile, given the coefficients used in the post-reform prediction are estimated from entirely pre-reform data.[21]

[19] We focus on fourth grade as third and fifth grade students were subject to a prior student-level accountability reform, and applying our procedure using them might conflate the effects of the two reforms.

[20] Here, we include a teaching approach that would target a specific point in the classroom distribution, as a matter of choosing an appropriate teaching style suited to the students in the class.

[21] For completeness, we compute a feasible pre-NCLB profile and show it in Appendix A.2. The resulting profile is

IV.B. Interpreting the Incentive Response

The incentive response profile in Figure 1 in hand, we now wish to provide an interpretation of the profile's estimated height for any given level of incentive strength. To inform that interpretation and also guide our modeling choices in the next section, we explore what *actions* are being taken, and by *which agents*?

Considering actions first, under school-level schemes such as NCLB, it is natural to think that schools would pull various levers at their disposal in order to try and attain the fixed accountability threshold. Here, the administrative data allow us to shed light on possible observable actions that schools might take, including: assigning the most marginal students to the best teachers (based on standard value-added measures); assigning the most marginal students to smaller classes; even making classrooms more homogeneous in terms of student ability. We can also cast indirect light on whether greater *effort* is given to marginal students, while acknowledging that we do not have direct effort measures in our data.

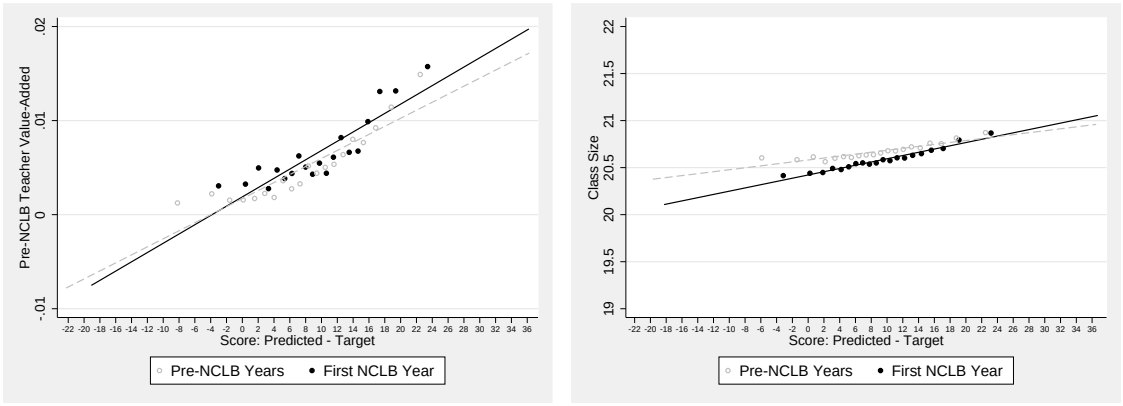
Taking these in turn, we examine changes in student-teacher *matching*, based on the notion that NCLB should lead higher ability teachers to be matched with more marginal students – in North Carolina, these are located low in the distribution, given the placement of the passing threshold at around only the 5th percentile. To investigate this, we plot the relationship between teacher ability and student incentive strength in Figure 2a, controlling for school-year-grade fixed effects and using pre-NCLB teacher value-added (VA) as our measure of teacher ability.[22] The positive relationship pre-reform, given by the dashed line, is indicative of positive assortative matching. Post-reform, rather than seeing the expected flattening of the relationship, it remains little changed (becoming steeper).

The fact that we do not see evidence of more able teachers being assigned to more marginal students suggests either that it is too costly to reassign teachers or that NCLB incentives may not override pre-existing *informal* incentives to focus on high-SES students. As we document below, informal incentives will play an important role in this North Carolina context.

Next we consider whether marginal students are being assigned to smaller classes in Figure 2b, which plots the relationship between class size and student incentive strength in both the pre-NCLB period and in the first year under NCLB. The relationship becomes slightly steeper after NCLB's introduction, indicating that marginal students may have been assigned to smaller classrooms in

much closer to zero throughout the distribution of incentive strength than the post-reform profile shown in Figure 1.

[22] Specifically, we use the standard jackknife Empirical Bayes estimator of value added, controlling for cubic polynomials of student prior scores and characteristics (see Kane and Staiger 2008, and Chetty, Friedman and Rockoff 2014).



(a) Teacher Ability (VA) (b) Class Size

FIGURE 2 – TEACHER ABILITY (VA) AND CLASS SIZE VERSUS STUDENT INCENTIVE STRENGTH

Notes: Panel (a) is a binned scatter plot of the relationship between teacher ability and student incentive strength. Panel (b) is a binned scatter plot of the relationship between class size and student incentive strength. These figures are constructed as follows: In both, the pre-NCLB period and the first year of NCLB, we calculate a predicted score for each student and then subtract off the known proficiency score target from this prediction – the horizontal axes in both panels measure the difference. In panel (a), the vertical axis measures teacher ability, which we estimate in the pre-NCLB period using the jackknife Empirical Bayes procedure. In panel (b), the vertical axis measures the number of students in each student's classroom – i.e., class size. In both the pre-NCLB period and the first year of NCLB, we residualize both the y-axis and x-axis variables with respect to school-grade-year fixed effects, adding back the unconditional mean of each variable to the residualized values to facilitate interpretation of the scale. Then we group students into 20 equal-width bins on the horizontal axis. Within each bin, we calculate the average of the y-axis and x-axis variables. The circles represent these bin-specific averages, while the straight lines represent the linear fits that are estimated using the underlying student-level data.

response. The scale on the vertical axis indicates, however, that average class sizes change by less than 0.1 students throughout the entire distribution of incentive strength – a magnitude unlikely to make any meaningful difference to students' test scores.[23] Further, there is no evidence of adjustments that could lead to a hump-shaped profile.

We also explore whether schools responded to NCLB by making classrooms more homogeneous. Creating classes post-NCLB in which students had similar academic preparedness could potentially make it easier for teachers to target instruction to a particular subset of students without necessarily increasing overall effort. To assess this possibility, we first provide an estimate of the maximal degree to which schools could make classrooms homogenous based on prior-year test scores and then compare actual classroom compositions against this benchmark.

A 'perfect sorting' benchmark can be established by reassigning students counterfactually to classrooms within a school-grade in a given year in order to maximize the between-classroom variance in prior scores.[24] For each school-grade, we then compare the actual fraction of the variance

[23]Indeed, we show this is the case directly, by recreating Figure 1 while controlling for classroom fixed effects below.

[24]Specifically, we arrange students in ascending order within each school-grade-year according to their prior-year mathematics score. We then fix the number of classrooms at the school-grade-year to be the actual number and proceed down the ranking to group students into these counterfactual classrooms while assigning each counterfactual class to have (approximately) the same number of students.

in prior scores that occurs between classrooms to the counterfactual fraction that would occur under the benchmark scenario. Doing so reveals that the distribution of students across classrooms falls well short of the perfect sorting benchmark. Specifically, the actual between-classroom variance accounts for only 6 percent of the total variance on average, yet it could account for as much as 78 percent if schools grouped students into classrooms based entirely on their prior scores.

Post-reform, if schools responded to NCLB by making classrooms more homogeneous, we would expect the observed between-classroom variance to account for a greater percentage of the maximum possible between-classroom variance in the 2002-03 academic year. Figure 3 plots the mean of this ratio over time (across all school-grades) relative to the baseline 1996-97 academic year, in which the average school-grade achieved a between-classroom variance that was 6.2 percent of the maximum possible variance. This fraction grows gradually over time, rising to 2 percentage points above the baseline-year value in 2004-05. There is no abrupt change in 2002-03, however, as the point estimate is not statistically different from the values in either of the two prior years (and only fractionally greater than any prior year). Further, the between-classroom variance at the average school in 2002-03 is only 8 percent of the maximum possible variance, indicating that schools tracking students into classrooms by prior ability was far from being a prominent margin of adjustment.

While Figures 2 and 3 provide evidence that schools did not engage in meaningful re-sorting of students in response to NCLB, one might still be concerned that the small documented change in the class sizes of marginal students or the gradual increase in classroom homogeneity (based on prior scores) could account for some of the distinctive pattern documented in Figure 1. One might also worry that unobserved determinants of tests scores that we cannot measure directly changed across classrooms in response to NCLB. To shed light on these issues, it is convenient to reproduce the NCLB response profile from Figure 1, and compare that with a corresponding profile constructed using only within-classroom variation.

Figure 4 shows the results. The near-perfect coincidence of the two profiles in the figure lends strong support to the view that differences *across* classrooms in teacher ability, class size, classroom homogeneity, or other unobservable factors do not drive the results. In addition, the use of classroom fixed effects in Figure 4 suggests that actions are taken mainly at the classroom level: specifically, teachers focus on directing their effort to more marginal students within-classroom. This evidence speaks to the agency question raised above. Also in line with such a *teacher* effort response, we can explore the relationship between teacher VA and classroom incentives (captured in a simple way by the proportion of marginal students, found within a given distance of the threshold). Looking *within-teacher*, as shown in Appendix Subsection A.3, we find that estimated teacher VA and classroom incentives under NCLB exhibit a positive relationship, while no such relationship is

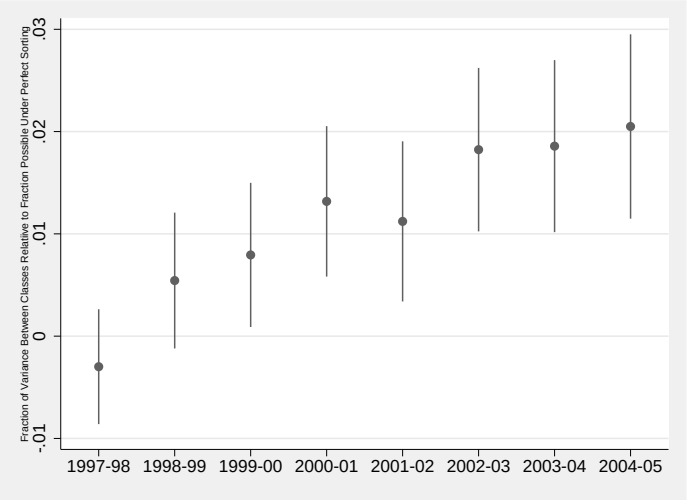


Figure 3 – Fraction of Total Variance in Student Prior Mathematics Scores Occurring Between Classrooms (Relative to Perfect Sorting Benchmark)

Notes: This figure shows the fraction of the total variance in students' prior-year standardized mathematics scores that occurs between classrooms relative to a counterfactual benchmark scenario in which schools sort students perfectly into classrooms based on prior-year scores. To construct the figure, we first sort students according to the 'perfect sorting' scenario (described in the main text). For each school-grade-year, we then determine the counterfactual within- and between-classroom variance in prior-year mathematics scores under this 'perfect sorting' scenario and compute both the actual fraction of the total variance that occurs between classrooms and the counterfactual fraction of the total variance that could occur between classrooms under the perfect sorting regime. We then divide the actual fraction by the counterfactual and regress this variable on year fixed effects and school-grade effects, treating the 1996-97 academic year as the baseline omitted category. The figure plots the estimated coefficients on the year fixed effects along with 95-percent confidence intervals. All estimates are relative to a baseline of 6.2 percent in the 1996-97 academic year.

apparent pre-reform. This is highly indicative of a within-teacher effort response.

In the model below, this evidence will lead us to treat teachers as the decision makers. At the same time, the model will capture *school*-level incentives, where success is defined in terms of each school attaining its target – a natural focus, given that incentives at the school level are the most common. School-level incentives have the practical advantage that aggregating to the school level guards against measurement error, yet they also give rise to possible free-riding on the part of individual teachers. We show that free riding does not appear to be a serious issue in our setting: specifically, comparing larger versus smaller schools, we do not find evidence that teacher effort is significantly lower as school size increases (see Appendix Subsection A.4). This evidence will lead us to abstract from free-riding concerns in the model that follows.

To summarize: based on the evidence we have assembled, the most plausible explanation for the inverted-U pattern in Figure 1 involves teachers adjusting discretionary effort in response to incentives. This targeting of what we term 'effort' to marginal students could take a variety of unobserved forms: teachers raising their energy levels and delivering material more efficiently to

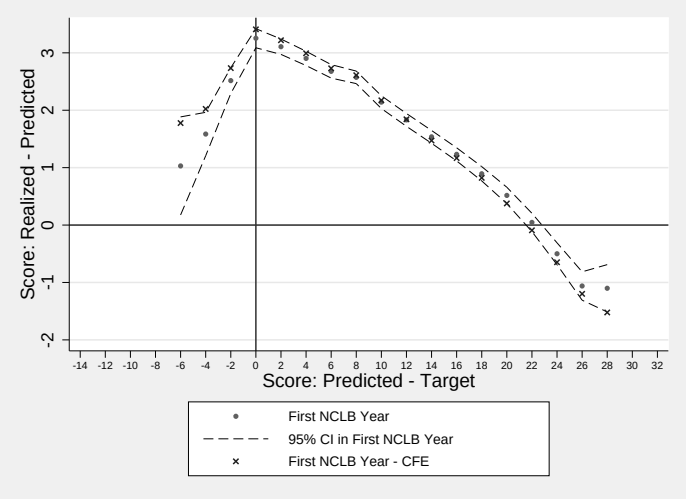


FIGURE 4 – STUDENT-SPECIFIC TARGETING

Notes: This figure presents the mean test score gains in the first year of NCLB from Figure 1, along with the 95-percent confidence intervals. It also shows the mean gains in that year, estimated using only within-classroom variation in test score gains and incentive strength. To construct the adjusted means using only within-classroom variation, we first regress (at the student level) gains above predicted scores on a mutually exclusive and exhaustive set of indicators for the bins on the incentive strength axis and classroom fixed effects. We then predict adjusted mean gains as the estimated coefficient on the indicator for each bin, without including the estimated classroom fixed effects in the prediction.

marginal students, or teaching some students more intensively 'to the test.'[25] While we do not have direct information about different actions taken by individual educators within the classroom, the post-reform profile takes on a shape that can be rationalized based on the known incentives. Further, in line with that explanation, we do not find evidence of changes in important non-effort inputs to education production: the inverted-U pattern is unchanged when controlling for teacher ability or class size (through the use of classroom fixed effects above), so the incentive response cannot be explained by school principals assigning marginal students to higher-ability teachers or smaller classes, nor to school principals making classrooms more homogeneous, which could otherwise allow for teachers to better target instruction to marginal students without increasing overall effort. The teacher-focused emphasis will be reflected in the model we develop next.

V. RATIONALIZING THE EFFORT FUNCTION: AN ESTIMABLE MODEL

We now wish to rationalize the effort function shown in Figure 1 in terms of underlying primitives of the effort-setting process, including those under the control of the policy maker. Our approach will be to develop a flexible effort-setting model that is informed by the evidence above and also

[25]The administrative data do not provide accurate information about the deployment of teacher aides, either across or within classrooms. The evidence above does not support the view that aides are being shifted *across* classrooms. To the extent that the reform leads teacher aides to be shifted to marginal students within-classroom, we will categorize that response as a component of 'teacher effort' – one that cannot be separately isolated in our setting.

relevant features of the North Carolina context. In the model, teachers choose effort optimally by balancing the benefits against the costs, providing – as we will argue – a natural lens for interpreting the estimated effort profile. In turn, the model will provide the essential foundation for the counterfactual analysis of incentive schemes that follows.

V.A. General Features of the Model

By way of overview, effort is the primary *action* taken by educators in the model in response to heightened accountability: the findings in the previous section align with there being a conscious effort response to the new incentives. In turn, we treat *agency* at the teacher level. The within-classroom evidence we have presented is consistent with a devolved process by which schools let teachers choose how much effort to devote to individual students in their classroom, leveraging their local information; here, of note, free-riding does not appear to be an issue.

Given our focus on teachers and the effort decisions they make, we develop a tractable model that can serve two purposes: (i) rationalize the effort function in a plausible way, and (ii) be taken to the data. The model's form will be guided by institutional features of our setting. The introduction of NCLB, which provides our key source of exogenous variation, affects the benefit side of the teacher effort choice problem. Consequently, this is where incentives will enter in our formulation. In contrast, there is no compelling reason to think that the reform influenced the cost of teacher effort; thus the form of the cost function should remain unchanged before and after, reflected in the flexible convex cost-of-effort structure we use.

The profile of the estimated effort function will serve as a discipline on our modeling choices. Here, two features are noteworthy: the asymmetric shape, especially elongated to the right, and the distinctive way the profile remains high above zero *well* to the right. While a standard effort-setting model can produce some degree of right-skewed asymmetry (the first feature), we will show that the significant elevation in the upper tail cannot be explained well by many models that might come to mind; those include models with complementarities on the production side, peer effects, and inaccurate teacher expectations, among other alternatives. In contrast, our chosen formulation with NCLB affecting the benefit side of the problem does a very good job of fitting the semi-parametric profile, as we will show.

Next, we will describe the main features of the model, alongside viable alternatives. (A detailed discussion of our modelling choices can be found in Appendix B.)

V.B. Model Specifics

The model has four main elements: a production technology, an accountability incentive scheme, a corresponding expected benefit function, and a cost of effort function. From these elements, we construct the teacher's objective, which will allow us to express the teacher's optimal effort choice in terms of key parameters, including features of the incentive scheme. The model we write down is estimable, by design, allowing those parameters to be recovered on the basis of an estimation routine presented in the next section.

Production Technology: We consider a technology that relates measured education output, labelled y, to various inputs, including 'effort' – the discretionary actions of educators that can increase output (as already described).[26] In our formulation, 'effort' will 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 (consistent with the evidence in the previous section). 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 – students are treated as passive (rather than active) learners. We summarize the exogenous inputs in a single measure, drawing on the discussion above – the *predicted score* for student i, $\hat{y}_i$. Increases in this measure will capture more favorable exogenous 'production' conditions.[27]

In practice, the underlying technology is known to be complex and is imperfectly understood. As an approximation to the true technology, we assume the following additive structure, consistent with the treatment effects apparatus above:

$$y_i = \hat{y}_i + e_i + \epsilon_i \,, \tag{2}$$

where y_i is student i's observed score, $\hat{y}_i$ captures all the exogenous inputs for student i, e_i is the teacher effort directed to student i post-reform, over and above baseline effort, and ϵ_i is a shock to test scores.[28] We assume that the error has a cumulative density function given by $F(\cdot)$, with mean 0 and variance σ^2.

[26] 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."

[27] Our goal, looking ahead, will be to predict the effects of alternative incentive schemes. With that in mind, we make minimal assumptions about the production technology, focusing on the role of the teacher effort input (rather than modeling the contributions of other educational inputs in detail).

[28] We omit time subscripts, as the model will be estimated using the impact of the NCLB incentive reform in 2002-03, rather than also relying on variation from subsequent years. 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.

Additivity is a minimal assumption, and one that is quite standard in the education literature. While various input interactions can be readily allowed, we note that in our setting, non-negligible complementarities (between student ability and teacher effort, for example) can be ruled out (see Appendix B.2). We also note the flexibility of the specification: $\hat{y}_i$ is a prediction using all prior information, and e_i is a semi-parametric data-driven function of incentives, drawing on the treatment effects estimation from the previous section.

Accountability Incentives: The accountability schemes we consider define a clear performance metric along with corresponding rewards or punishments. 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 for. 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. We will interpret it more broadly, allowing it to 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). In the absence of further information, we suppose all teachers share the same benefit from attaining target level l.

In the context of our study (especially, the actual incentives used to estimate the model parameters), the NCLB incentives consist of one explicit fixed target level 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. A simpler model that only allows for teachers to respond to the NCLB target can be ruled out (see Appendix B for the detailed argument).

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 – the evidence in the previous section 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 to $y_l^T + d_l$.[29] The shifter parameters are homogeneous with respect to teachers and schools. This is not simply for tractability: the evidence indicates that the semi-parametric effort function is invariant to various teacher and school characteristics – teacher ability and class size (already shown), as well as the prior preparation of a teacher's students relative to the school distribution (see Appendix Subsection A.5). The benefit is then given by:

$$\begin{aligned} B(e_1, \ldots, e_{N_c}) &= b_M \sum_{i=1}^{N_c} Pr[y_i > y_M^T + d_M] + b_H \sum_{i=1}^{N_c} Pr[y_i > y_H^T + d_H] \\ &= b_M \sum_{i=1}^{N_c} \Big[1 - F(y_M^T + d_M - \hat{y}_i - e_i)\Big] \\ &\quad + b_H \sum_{i=1}^{N_c} \Big[1 - F(y_H^T + d_H - \hat{y}_i - e_i)\Big], \end{aligned} \tag{3}$$

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

Cost of Exerting Effort: Teacher c faces a convex cost of effort that has the following flexible form:

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

The parameter ψ 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.[30] 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 complex and 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. Doing so will allow us to explore the counterfactual implications of incentive design. We focus on the effort decisions of teacher c, allowing each student to receive student-

[29] Such overshooting offers an additional degree of freedom when matching the effort profile. Indeed, it is necessary in order to rationalize maximal effort being directed toward students with predicted scores that equal the target (consistent with the effort pattern recovered in the previous section). This will be made clear when we describe the conditions for optimal effort-setting in the next subsection.

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

specific effort e_i. This is reasonable given the clear evidence from the previous section that the teacher effort response is driven almost entirely by within- rather than across-classroom variation in production conditions (summarized by $\hat{y}$). Because formal incentives under NCLB apply to schools as a whole, it may seem surprising that effort decisions would be taken by individual teachers; a natural interpretation is that school principals delegate 'local' decisions to teachers in service of school-level objectives, in line with their likely informational advantage.[31]

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:

$$U_c = B(e_1, \ldots, e_{N_c}) - C(e_1, \ldots, e_{N_c}), \tag{5}$$

where the explicit functions are given by equations (3) and (4).

V.C. 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 (5) 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:

$$\begin{aligned} \frac{b_M}{\psi} f\left(y_M^T + d_M - \hat{y}_i - e_i^*\right) \; + \frac{b_H}{\psi} f\left(y_H^T + d_H - \hat{y}_i - e_i^*\right) & \\ = \left[e_i^* + \theta \sum_{j=1}^{N_c} e_j^*\right], \; \forall \; i = 1, \ldots, N_c, & \end{aligned} \tag{6}$$

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 $\hat{\mathbf{y}}_c \equiv \{\hat{y}_j\}_{j \neq i}$.[32]

[31] 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.

[32] 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 conditioning variable, $\hat{\mathbf{y}}_c$.

The Solution – Intuition: As specified, optimal effort e^* does not have a closed-form solution.[33] Still, the model structure allows us to provide intuition regarding the way optimal effort is determined. 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.[34]

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

$$\frac{b_M}{\psi} f\left(d_M - \pi_i - e_i^*\right) \; = \left[e_i^* + \theta \sum_{j=1}^{N_c} e_j^* \right], \tag{7}$$

using equation (6) 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 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 $\hat{y}_i$) increases, it follows that the marginal benefit curve shifts leftward.

On the cost side, 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 patterns documented in the previous section. The asymmetry we find in the effort function, skewed to the right, is then 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), a feature that affords a better match with the effort profile.[35]

Now consider the case in which the higher target also influences effort setting, alongside the NCLB target. The optimal effort profile will now be the vertical summation of the effort profiles implied by each of the two separate targets, as reflected in equation (6). The higher target will give rise to a second peak in the effort profile; this will provide a fruitful way to explain the estimated

[33] An iterative process is required to determine optimal effort as the derivatives of the parameters in the nonlinear first-order condition depend both on the effort level of interest and the parameters themselves.

[34] Looking ahead to the estimation section, this will align with Assumption 4 below.

[35] 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 $\theta \sum_{j \neq i}^{N_c} e_j^*$. 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. When $\theta > 0$ and average classroom effort is positive, the model permits solutions to equation (7) in which optimal effort is negative.

shape of the effort function, considered below.

Properties: Building on this discussion, 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. To do so, write optimal effort for student i (determined as the solution to equation (6)) 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; Appendix C provides a detailed analysis.

Several comparative static results are worth noting. First, the *spread* of the effort profile is increasing in σ^2.[36] This relationship is illustrated 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.[37] 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.[38] 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 the peak of the marginal benefit curve, which in turn affects the height of the peak in the optimal effort profile. These properties will be useful for understanding model identification in what follows.

VI. Model Estimation and Identification

This section describes the estimation of the model, and considers 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.

[36] The reason is that the marginal benefit curve broadens as σ^2 increases, slowing the rate at which effort declines away from its peak as it shifts against the marginal cost curve (where the intersection determines the effort solution).

[37] 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.

[38] This occurs since each parameter affects the associated peak of the marginal benefit curve, which affects the maximum of the optimal effort profile.

Using the production technology in equation (2) and the first-order condition implicitly defining optimal effort in equation (6), the test score for any student i can be written

$$y_i = \hat{y}_i + e^* \left(\beta; \hat{y}_i, \mathbf{y^T}, \hat{\mathbf{y}}_c\right) + \epsilon_i \,. \tag{8}$$

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 (8) and make a further distributional assumption:

Assumption: 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 observed and predicted score data, as well as the effort level[39] implied by the model's parameters:

$$\begin{aligned} L_i(\beta) &= f\left(\epsilon_i \mid \frac{b_M}{\psi}, \frac{b_H}{\psi}, d_M, d_H, \theta, \sigma^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{1}{2\sigma^2} \cdot \left(y_i - \hat{y}_i - e^*(\beta; \hat{y}_i, \mathbf{y^T}, \hat{\mathbf{y}}_c)\right)^2\right\} . \end{aligned} \tag{9}$$

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:

$$\begin{aligned} \ell(\beta) &\equiv \sum_{i=1}^{N} \log L_i(\beta) \\ &= -\frac{N}{2} \cdot \log(2\pi) - \frac{N}{2} \cdot \log \sigma^2 - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{N} \left(y_i - \hat{y}_i - e^*(\beta; \hat{y}_i, \mathbf{y^T}, \hat{\mathbf{y}}_c)\right)^2 . \end{aligned} \tag{10}$$

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 vector and solves equation (6) 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 test score for each student as $\hat{y}_i + e^* \left(\beta; \hat{y}_i, \mathbf{y^T}, \hat{\mathbf{y}}_c\right)$, which allows us to evaluate the likelihood in equation (10) in the

[39] It is worth noting that effort for a given student i is not simply read off from the corresponding height of the estimated profile, yielding $e_i = e^*(\pi)$. Our implementation is necessarily more complex, given the optimal effort applied to a particular student as a solution to the teacher effort problem depends on the vector of predicted scores of *all* students in the corresponding classroom (following the form of the cost function). Thus, while two students in different classrooms may have the same predicted score, the optimal effort levels applied to them can differ, depending on the classroom distributions of exogenous student characteristics.

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

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_M^T) in fourth grade; similarly, the target required for 'superior' performance (y_H^T) 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 are presented in Table H.1. The table makes clear this sample looks quite similar to the full sample of students (in Table 1) in terms of demographic variables, although mean test scores are slightly higher in the former.

VI.B. Identification of Parameters

We now discuss the identification of each of the main parameters of interest.

The σ^2 parameter: This parameter captures the variability in effort from the model's predictions. It 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. 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.

The θ parameter: This governs the within-classroom tradeoffs in effort that teachers must make across their students. The evidence in Section V.C 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 to any student i is an increasing function of the effort directed to any other student. 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.[44]

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

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

[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] A positive value of θ implies a negative horizontal intercept for the MC curve. Thus, there will be two critical values of $\hat{y}$ (one low and one high) for which the intersection of the MB and MC curves (and thus implied effort)

The $\frac{b}{\psi}$ 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.C). 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.[45] 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. 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.[46] 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.

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'

will turn negative.

[45] We expand the argument to allow for possible interactions in Appendix C.

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

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.[47] 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 ($d_M = 3.19$).

Table 2 – Parameter Estimates

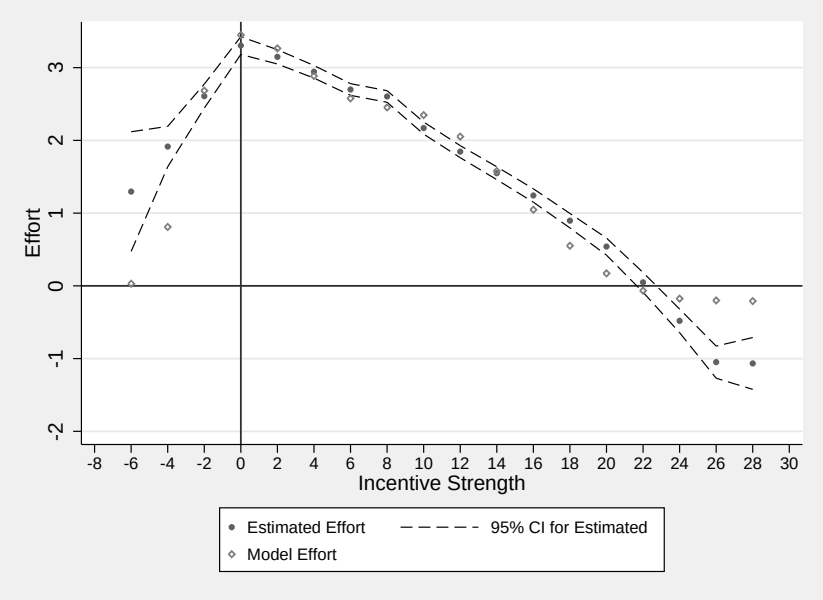
Parameter	Estimate
$\frac{b_M}{\psi}$	36.297***
	(0.6234)
$\frac{b_H}{\psi}$	24.001***
	(0.5450)
d_M	3.1945***
	(0.0647)
d_H	1.634***
	(0.0950)
θ	0.0075***
	(0.0010)
σ^2	15.702***
	(0.0084)
N	89,271

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

VII.B. Model Fit

In terms of model fit, we start by plotting the 2002-03 data from Figure 1 along with the effort predicted by the model. We use the model to generate effort 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 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.

[47] 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 (higher $\hat{y}$) students also gained following NCLB's introduction.



Notes: This figure presents the profile of the estimated effort function in 2002-03 from Figure 1 along with the 95 percent confidence intervals for that function and the binned means of model-implied effort.

FIGURE 5 – INVERTED-U RESPONSE TO NCLB AND MODEL FIT

Next, we can use equation (8) 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 (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 a reasonably accurate way.

Figures E.1 and E.2 in Appendix E 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.

VIII. Counterfactual Framework

In this section, we set out our counterfactual approach, based on a framework that combines the plausible structure of the model with the estimates from the previous section. Using this, we can

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

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

explore a variety of alternative incentives 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 are able to 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[48] – as well as the remainder, constituting the majority.

We present the simulation framework next. Then we describe the set of proficiency targets considered using the framework and a cost-equating procedure to ensure comparability, before turning to the counterfactual results themselves in the following section.

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

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 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. 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 (5).[49]

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. Students' predicted scores in the absence of accountability incentives are held fixed across all of our counterfactual simulations, and we keep the model's parameters 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 ($b_i \equiv w_i \cdot b_M$) via 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.[50] Optimal effort for student i is given by $e_{it}^* = e^*\big(w_i, y_{it}^T; \widehat{\beta}, \hat{y}_{it}, \hat{\mathbf{y}}_c\big)$.

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

[50] Here we follow the same computational procedure as in the model above (described in full in Appendix D): that is, we solve N_c first-order conditions in each classroom simultaneously to recover the full distribution of effort.

Technology and Counterfactual Output: With the counterfactual effort vector in hand, we then use the technology in (8) 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:

$$y_{it} = \hat{y}_{it} + e^*\big(w_i, y_{it}^T; \widehat{\beta}, \hat{y}_{it}, \hat{\mathbf{y}}_c\big) + \epsilon_{it}, \tag{11}$$

where $\hat{y}_{it}$ is the student's predicted score based on all prior information,[51] $e^*\big(w_i, y_{it}^T; \widehat{\beta}, \hat{y}_{it}, \hat{\mathbf{y}}_c\big)$ 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 (11) can then be used to recover the associated test score distribution across all students.

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 involve targets that are the same for all students – for example, students 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 standard intuition, 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 all), spanning the full predicted score distribution.[52]

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

[52] See Appendix F.1.

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, a higher weight (in the form of a higher student bonus) is given 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 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 $y_{it}^T = \delta + \alpha y_{i,t-1}$, the relevant threshold benefit rule written $b_i \cdot 1(\, y_{it} \geq y_{it}^T)$, as before. The parameter δ influences the mean of the incentive strength $(\hat{y} - y^T)$ distribution, while α governs the variance of that distribution.[53]

We explore the effects of different VA targets on outcomes by varying the target parameters systematically, as follows: each value-added target can be linked precisely to a corresponding fixed target, noting that a fixed target is a special case of a VA target in our formulation, 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.[54] 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.[55]

VIII.C. Cost Equating

For comparability, we place all the counterfactual incentive schemes under consideration on a common footing. Specifically, we ensure that every target regime results in the same cost, changing the bonus payment until we achieve cost-equivalence across regimes. Having ensured cost-equivalence

[53] 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.

[54] 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.

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

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

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.[56] The essence of the cost-equating procedure in this case is as follows: given teachers' optimal effort choices are influenced by the incentive parameters in the effort-setting first-order condition, changing the actual bonus payment b_M counterfactually will influence passing probabilities, and thus the implied cost of the resulting scheme. While b_M is not separately identified in our estimation framework, as noted, 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 then pick the value of k that equates the cost to the actual cost (equivalent to the passing rate) under the prevailing NCLB target.

The cost-equating procedure when bonuses are student-specific is somewhat more involved. Appendix F.4 gives a fuller description of this case, and the cost-equating procedures we follow more generally.

IX. Counterfactual Results

This section presents the main results of our counterfactual policy 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, as these capture notions of efficiency and inequality: other informative measures are easily computed.

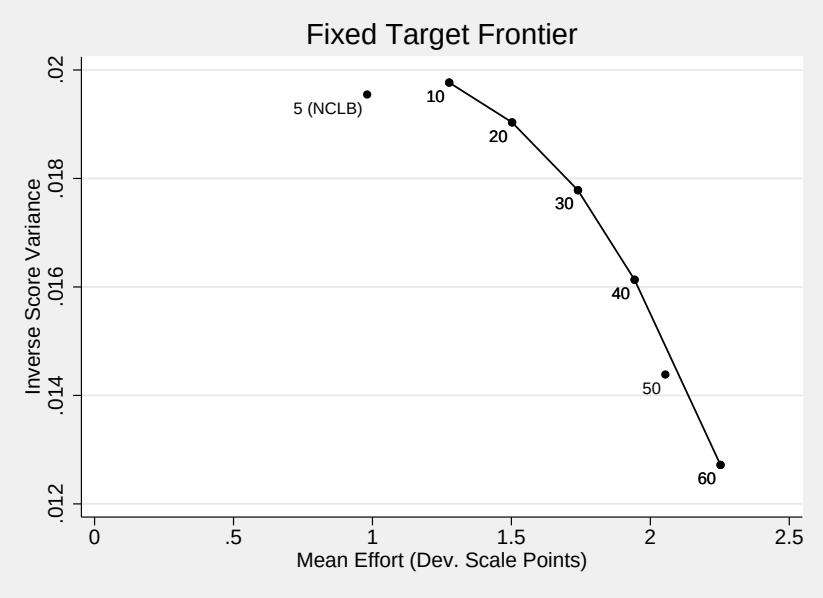
IX.A. Fixed Targets with Homogeneous Bonus Payments

Schemes featuring fixed targets that do not alter the bonus according to student type are easily 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

[56] 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.

effort and a measure of spread – the inverse of the test score variance[57] – 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. In Figure 6, we plot the resulting 'mean effort-inverse variance' points to trace out the frontier. 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.



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 6 – FIXED FRONTIER WITH HOMOGENEOUS BONUS PAYMENT

The frontier shows a 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 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

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

students, thereby raising mean effort.[58] 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, the explanation for the downward slope becomes somewhat more subtle. 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. 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). Thus 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.

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.

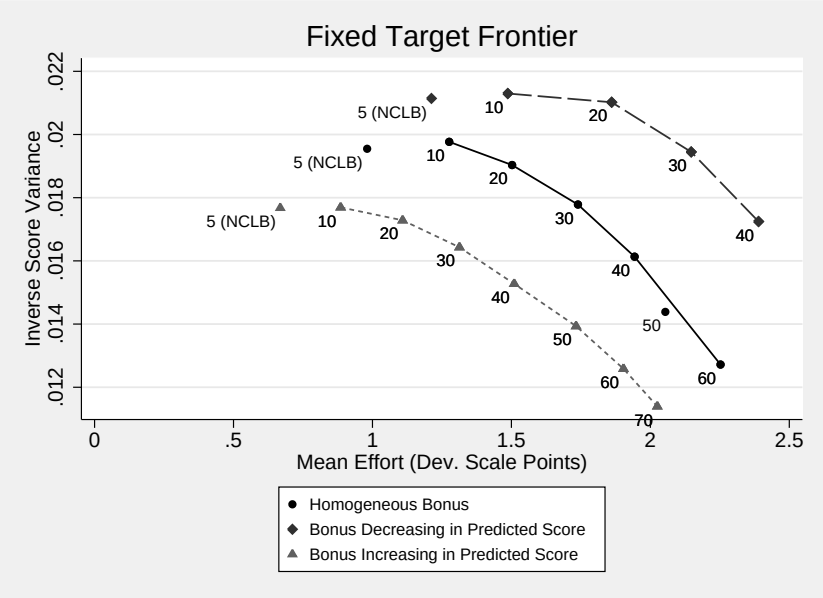
In particular, Figure 7 shows that the scheme placing more weight on low-performing students dominates the homogeneous bonus payment regime, which in turn dominates the scheme that places more weight on high-performing students. In terms of magnitudes, suppose we hold the proficiency target fixed at 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]

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 NCLB) and switching to this regime from the 'constant bonus payments' regime.[60] The new regime

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

[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.

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



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 6. 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 7 – FIXED FRONTIERS WITH HETEROGENEOUS BONUS PAYMENTS

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 would 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. This is because we must increase the bonus payment 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).

It is clear from Figure 7 that the scheme that assigns more weight to high-performing students is dominated by both other regimes. The forces that lead to the inward shift of the frontier when switching to this regime follow a similar logic, now in reverse.[61]

[61] In this case, the essence is as follows: 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

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 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 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.

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. (See in Appendix G.3.)

IX.C. Value-Added Targets

We now explore the properties of value-added (VA) targets. As we will see, VA targets provide the policy maker with an additional lever, compared to fixed targets: by incorporating student-specific prior information, they provide a means to adjust the variance of incentive strength across students. In turn, this additional lever allows the policy maker to promote the equalization of effort across across students, rather than teachers focusing on marginal students – consistent with students receiving the same level of investment while attending the same school.

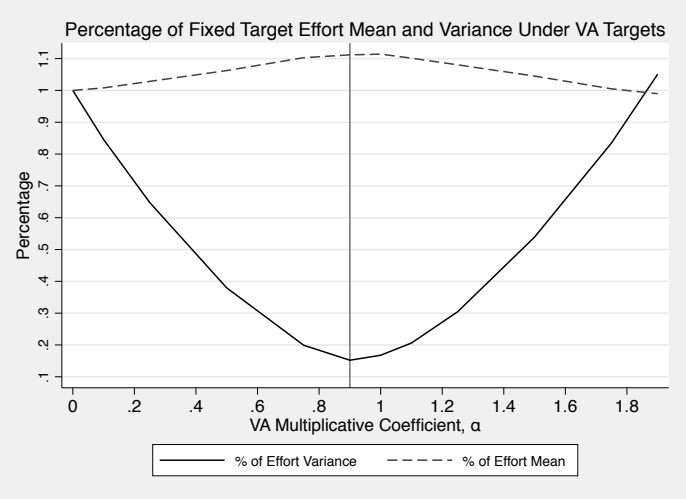
In what follows, we exploit the linkage between fixed and value-added targets described in the previous section, whereby each VA target shares the same incentive strength mean (across all students) as a given fixed target but can have a different variance.[62] Our interest centers on how outcomes change 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, given by $cov(\hat{y}_t, y_{t-1})/var(y_{t-1})$, is straightforward to derive.[63] 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.

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 8 shows that, compared against 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 8 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^*.

[62] 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.

[63] 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.)

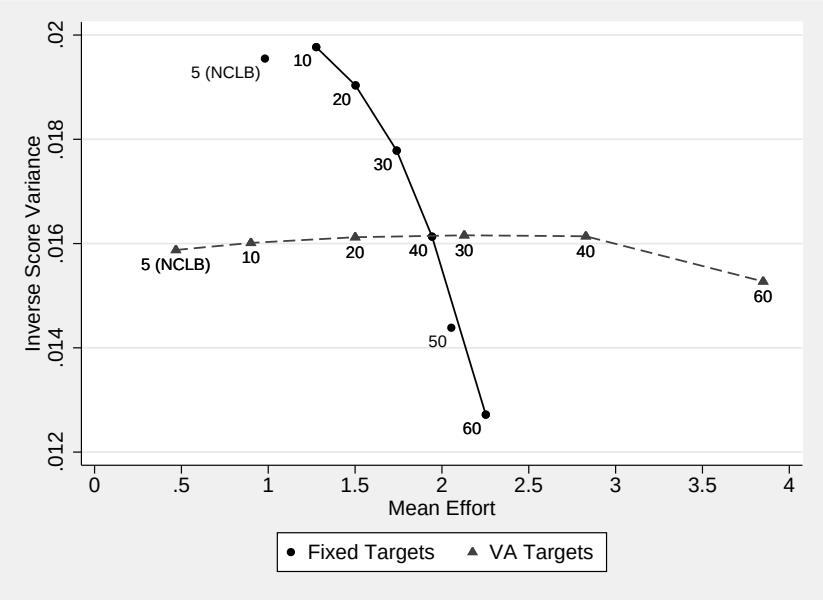


Notes: This figure shows the percentages of the fixed target effort mean and variance that are attained by VA targets with different multiplicative coefficients, α. The vertical line indicates the effort-variance-minimizing α, namely α^*, equal to 0.937 in our data. The dashed line shows the average 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 8 – PERCENTAGE OF FIXED TARGET MEAN AND VARIANCE ACHIEVED BY VA TARGETS

Next, Figure 9 shows the mean and variance properties of VA targets in terms of effort (alongside the fixed frontier from Figure 6, for reference), while setting $\alpha = \alpha^*$ for all VA targets. Because of the incentive to equalize effort, 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 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. Intuitively, 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 fortieth 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



Notes: The solid line plots the frontier from Figure 6. 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 9 – FIXED AND VA TARGET FRONTIERS

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 set out a new approach for studying the design of incentives, applied to an education context. Our approach is built around the estimated relationship between accountability incentives and teacher effort. Here, we showed how features of the North Carolina context (in particular, the exogenous incentive variation associated with the introduction of a prominent accountability reform) could be used to identify the effort response of teachers based on changes in test scores. Our method for doing so 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) – valuable given that effort is typically unobserved and thus difficult to pin down.

We then proposed a structural procedure based on a flexible model of effort setting, which we used to identify the primitives underlying the teacher effort response. Estimates of the model showed that within-classroom tradeoffs in effort across students are important, and that teachers boosted effort following NCLB's introduction.

The model and estimates formed the basis of a counterfactual framework at the heart of the paper for measuring the performance of different feasible 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, generating policy-relevant insights. In particular, we compared 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 policy analysis, 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.

Beyond the current analysis, our counterfactual approach provides a valuable policy design tool at a time when states are re-visiting education incentives. By offering insight into the distributional consequences of education accountability policies, it clarifies how education reforms can be used 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.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2007. "Incentives for Managers and Inequality among Workers: Evidence from a Firm-Level Experiment." *Quarterly Journal of Economics*, 122(2): 729-773.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2010. "Social Incentives in the Workplace." *The Review of Economic Studies*, 77(2): 417-458.

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.

Carnoy, Martin and Susanna Loeb. 2002. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." *Education Evaluation and Policy Analysis*, 24(4): 305-331.

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.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness" *Journal of Policy Analysis and Management*, 23(2): 251-271.

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, and Christopher Jencks 2016. "School Accountability, Postsecondary Attainment and Earnings." *Review of Economics and Statistics*, 98(5): 848-862.

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 Lawrence S. Getzler. 2006. 'Accountability, Ability and Disability: Gaming the System?." *Improving School Accountability (Advances in Applied Microeconomics)*, 14: 35-49.

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

Figlio, David and Joshua Winicki. 2005. 'Food for thought: the effects of school accountability plans on school nutrition." *Journal of Public Economics*, 89(2-3): 381-394.

Hanushek, Eric A. and Margaret E. Raymond 2005. 'Does school accountability lead to improved student performance?." *Journal of Policy Analysis and Management*, 24(2): 297-327.

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.

Kane, Thomas J. and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." National Bureau of Economic Research Working Paper 14607.

Kory Kroft, Yao Luo, Magne Mogstad, and Bradley Setzler. 2021. "Imperfect Competition and Rents in Labor and Product Markets: The Case of the Construction Industry."

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

Thibaut Lamadon, Magne Mogstad, and Bradley Setzler. 2021. "Imperfect Competition, Compensating Differentials and Rent Sharing in the U.S. Labor Market."

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. 2019. "Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievements." *Journal of Labor Economics*, 37(3): 621-662.

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

Macartney, Hugh, Robert McMillan, and Uros Petronijevic. 2020. "Teacher Value-Added and Economic Agency." Revised version of NBER Working Paper 24747.

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.

Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature*, 37(1): 7-63.

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

Appendices

A. Incentive Responses: Further Evidence

This appendix presents evidence to supplement the discussion in Section IV and Section V.

A.1. Prediction Accuracy

As a gauge of the accuracy of our prediction algorithm, described in Section IV, we present the following figure.

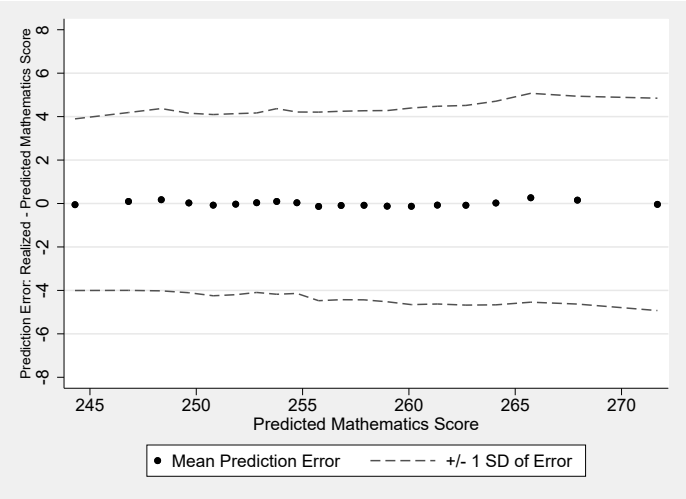


Figure A.1 – Prediction Error versus Predicted Mathematics Score

Notes: This figure plots the error in predicted mathematics scores along the y-axis, defined as the difference between the realized and predicted score, against the predicted score on the x-axis. The predicted score is the fitted value from a student-level regression of current mathematics scores on cubic functions of prior mathematics and reading scores, along with indicator variables for student race, gender, free-lunch status, English proficiency status, and disability status. The R^2 from this regression is 0.727. Along the x-axis, we group students into 20 equal-sized bins (with almost exactly 4,000 observations per bin) and calculate the average of the y- and x-axis variables within each bin. The circles represent these averages. Across all student observations within a given bin, we also calculate the standard deviation of the prediction error. The dashed lines indicate plus and minus one standard deviation of the mean prediction error in each bin.

What is striking, as noted in the main text, is the way the mean deviation of the predicted from the actual score is very close to zero for each bin in the figure, throughout the predicted score distribution.

A.2. Pre-Reform Treatment Effect Profile

For completeness, we compute a pre-NCLB profile and place it alongside the inverted-U profile shown in the main text. The profile is for 1999-00, predicted using data from 1998-99 and 1997-98.[64]

What is apparent from the figure is just how much closer to zero the pre-reform profile is than the post-reform profile shown in Figure 1. There is no obvious explanation for the estimated 'bumps' that we

[64] A change in the developmental scale of test scores in 2000-01 means that a profile for 2001-02 (the year immediately preceding the introduction of NCLB) or 2000-01 cannot be constructed.

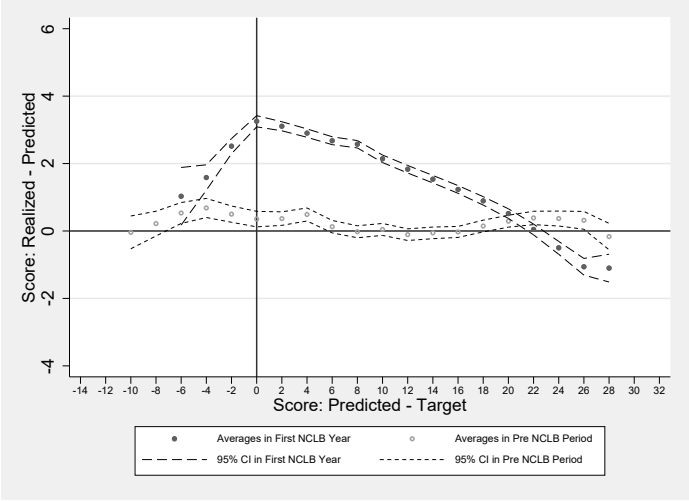


FIGURE A.2 – EFFORT RESPONSES & PLACEBO

can discern in the pre-reform profile. At the same time, there is no reason to expect that this 'placebo' profile should be entirely flat anyway: *any* factor in the pre-reform years that operates at a particular range in the incentive strength distribution would affect the estimated profile. We also note that any such factors are already accounted for in the algorithm used to construct the post-reform profile.

A.3. Within-Teacher Performance Improvements

We explore whether stronger NCLB incentives caused greater *within-teacher* performance improvements, relevant to the agency issue. We do so by relating the difference in teacher-year fixed effects for a given teacher (after vs. directly before the reform) to the fraction of marginal students taught by each teacher in 2002-03, the year NCLB came into effect. In this context, the fraction of marginal students ($m_{j,02-03}$) is defined very simply as the proportion of students with a value of the incentive measure ($\hat{y} - y^T$) between -4 and 4.[65] We can also carry out an informative 'placebo' comparison using successive years prior to the reform.

Specifically, we estimate the following equation at the teacher level:

$$\hat{q}_{j02-03} - \hat{q}_{j01-02} = \alpha + \chi m_{j02-03} + g(\hat{q}_{j01-02}) + \zeta_{j02-03}, \tag{12}$$

where $\hat{q}_{jt}$ on the LHS is defined as the estimated teacher-year fixed effect obtained from a value-added regression of student test scores on demographic controls and teacher-year indicators in year t. The main parameter of interest is χ, reflecting any relationship between NCLB incentives and within-teacher performance improvements, and $g(\hat{q}_{j01-02})$ is a cubic function of 2001-02 teacher-year VA, which we include to account for mean reversion in teacher performance.[66]

[65] Other thresholds for defining a student as marginal yield similar results.

[66] Within-teacher fluctuation in performance could be driven by mean reversion when, for example, teachers with high fractions of marginal students 2002-03 were 'unlucky' in 2001-02 and had performed unusually poorly in that year. In that case, we would expect their performance to improve mechanically from one year to the next, independent of the new NCLB performance incentives. We account for any such mechanical relationship between lagged VA and

Figure A.3 shows a binned scatter plot of the partial relationship between the performance improvement in 2002-03 and $m_{j,02-03}$ along with the associated linear fit (given by the solid line) from the underlying teacher-level data; see Macartney *et al.* (2021) for a fuller treatment.

Within-teacher performance improvements are clearly increasing in the fraction of marginal students in the classroom in 2002-03, shown by the upward-sloping line. Because the specification in equation (12) removes any effect of (fixed) teacher ability, it is unlikely that differential sorting of students to teachers based on ability can explain our results. A pooled regression of all pre-NCLB years (with transitions from year $t-1$ to t) is used as a placebo control, showing a relatively flat relationship (given by the dashed line), and further supporting the claim that the estimated 2002-03 relationship reflects NCLB effort incentives.

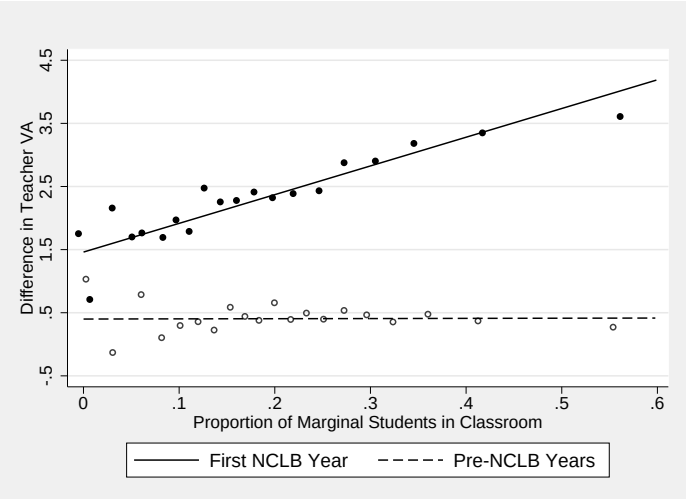


FIGURE A.3 – CHANGE IN VA VS. PROPORTION MARGINAL

Notes: This figure plots the change in teacher-year value-added for a given teacher between year t and $t-1$ as a function of the proportion of students in her year-t class who are marginal, comparing pairs of years before and after NCLB's introduction. To construct the figure, we first define a student as marginal if the difference between her predicted score and the NCLB proficiency target (the x-axis variable in Figure 1) is between -4 and 4, and then we define m_{jt} as the proportion of marginal students in the classroom of teacher j in year t. We also define $\hat{q}_{jt}$ as the estimated teacher-year fixed effect from a value-added regression of student test scores on demographic controls and a teacher-year indicator for year t. In both the first year of NCLB and in the pre-NCLB period, we then residualize m_{jt} with respect to a cubic function of $\hat{q}_{jt-1}$ and, in the pre-NCLB years, year fixed effects. We then add back the unconditional mean of m_{jt} to the residualized values to facilitate interpretation of the scale, and plot this variable on the x-axis. The y-axis measures the change in the estimated teacher-year fixed effects for each teacher between year t and $t-1$, given by $\hat{q}_{jt} - \hat{q}_{jt-1}$. On the x-axis, we group teacher-year observations into 20 equal-sized bins and calculate the average of the y- and x-axis variables within each bin. The circles and dots represent these averages. The lines represent the associated linear fits, estimated on the underlying teacher-year data.

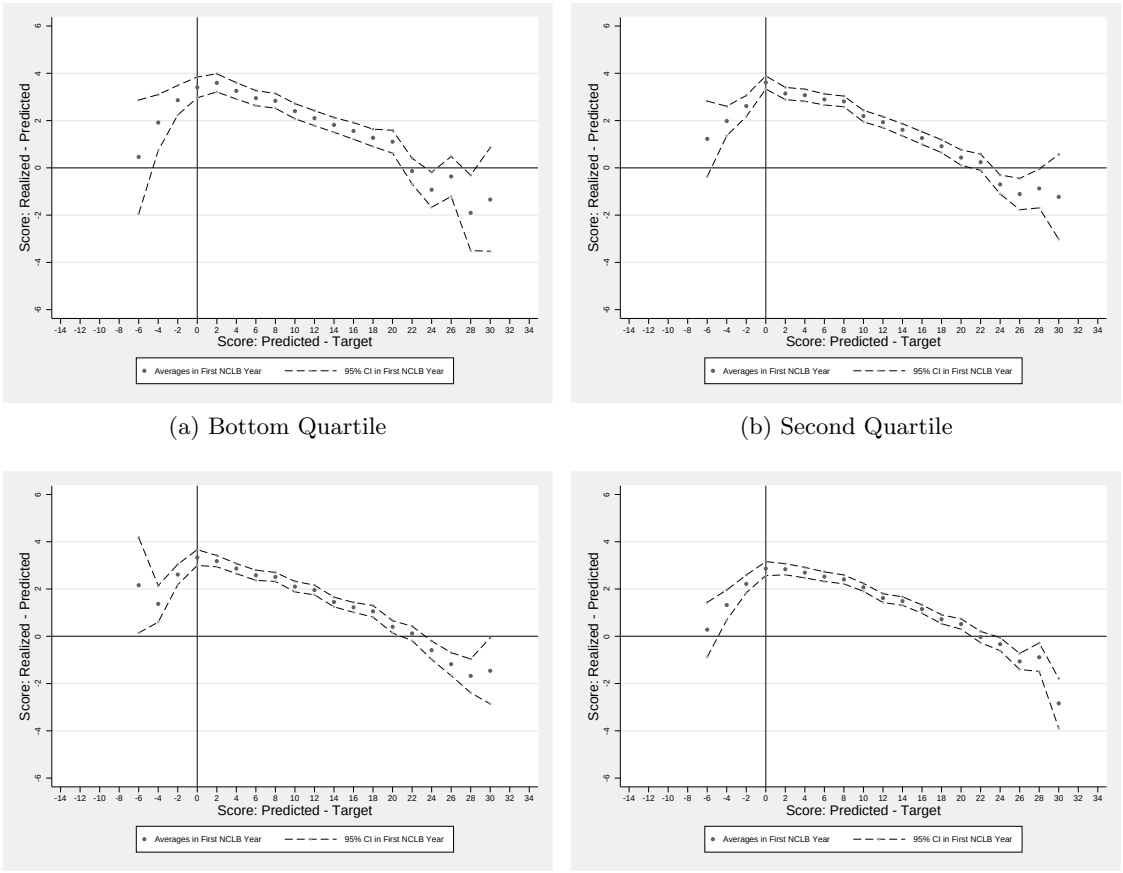
A.4. Free-Riding Concerns

In this subsection, we show that teacher free-riding is not a central driver of differences in the effort response across schools, helping to motivate the teacher-focused model in Section V.

Figure A.4 plots the same relationship as Figure 1, but it does so separately for schools of varying size. If free riding were an important factor influencing effort provision in our context, we would expect to see a

performance improvement in a flexible way (with the cubic function of lagged VA), thereby identifying the effect of NCLB incentives conditional on that relationship.

much compressed effort profile in the largest schools, where free-riding should be most severe, relative to the effort profile that prevails in the smallest schools, where there are fewer teachers to share in effort provision. The patterns in Figure A.4 across school quartiles suggest that free-riding concerns are not first order.



(a) Bottom Quartile

(b) Second Quartile

(c) Third Quartile

(d) Top Quartile

FIGURE A.4 – NCLB EFFORT RESPONSE BY QUARTILE OF SCHOOL SIZE

Notes: This figure plots the same relationship as Figure 1, but it does so separately for schools in different quartiles based on size. The y- and x-axis variables are measured and constructed as in Figure 1. We then further divide schools into quartiles of the school size (total student enrollment) distribution in the first year of NCLB and plot separate relationships between test score gains and incentive strength among schools in each quartile, panel (a) plotting the relationship for schools in the bottom quartile (the smallest schools), and so on (second, third and fourth quartiles in panel (b), (c), and (d).

A.5. Response to Target, Not Position in School-Specific Distribution

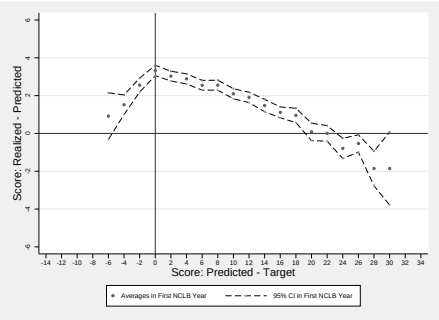
Our maintained hypothesis is that effort is responsive to the incentive measure (π) that we have constructed. As an alternative, effort might vary with respect to a student's *relative* position in the predicted score ($\hat{y}$) distribution within her school. For example, it is possible that educators 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.[67] 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 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 grouping them on the basis of which quartile (in terms of that predicted pass rate) they are in.[68] If schools responded to NCLB by tailoring effort to a particular part of the ability (i.e., predicted score) 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.

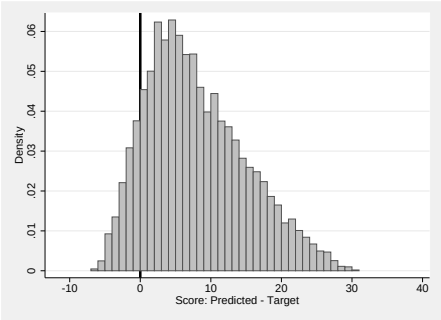
Figure A.5 plots the effort responses and incentive strength densities separately for schools in each quartile of the school-level *ex-ante* predicted pass rate. As one moves up the quartiles, the incentive distribution shifts rightward (as shown in the panels on the right), implying that a student with 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 in each of the quartiles, indicating that it is the students who are most marginal with respect to the NCLB threshold who receive most attention. 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.

[67] Such a response is in the spirit of Duflo, Dupas, and Kremer (2011), who set out a model in which teachers choose a particular method of teaching 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.

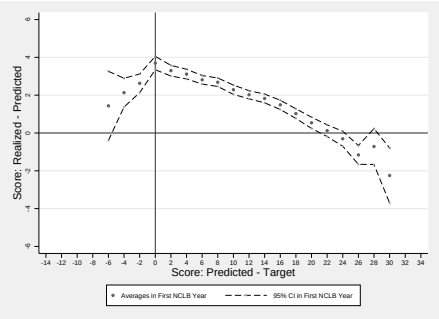
[68] Recall that a student is predicted to pass when $\pi = \hat{y} - y^T > 0$.



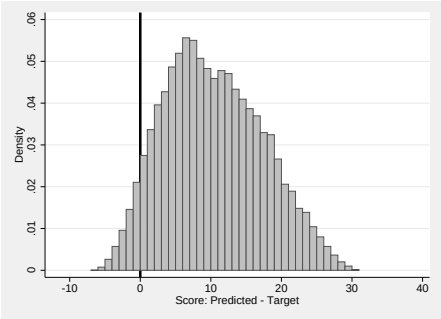
(a) Effort in Q1 Pass-Rate Schools



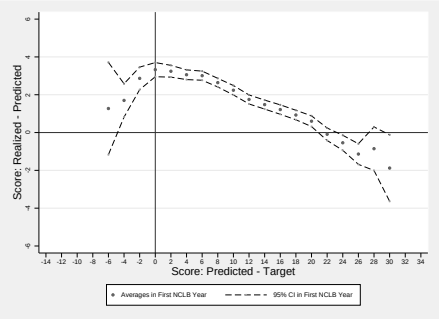
(b) π Density in Q1 Pass-Rate Schools



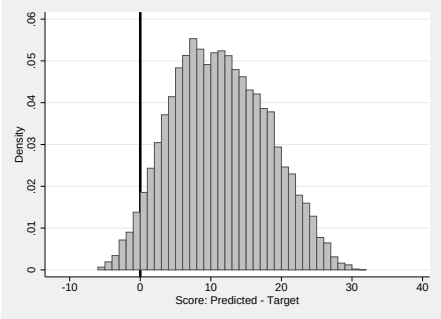
(c) Effort in Q2 Pass-Rate Schools



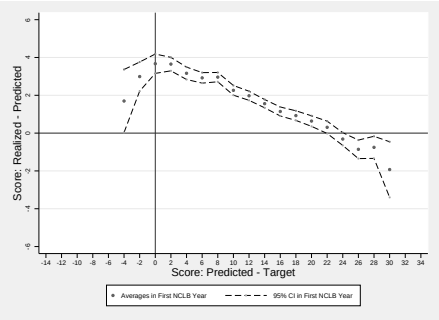
(d) π Density in Q2 Pass-Rate Schools



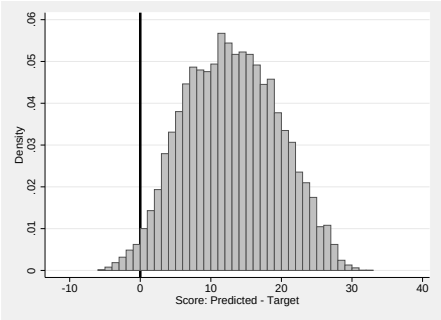
(e) Effort in Q3 Pass-Rate Schools



(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.5 – RESPONDING TO INCENTIVE STRENGTH π RATHER THAN THE RELATIVE POSITION OF $\hat{y}$

B. Modeling Choices

Formal NCLB incentives only apply to the attainment of a single target – the middle proficiency target under North Carolina's pre-existing assessment system. Our model in Section V provides additional flexibility, reflecting institutional arrangements in North Carolina. It allow teachers to respond to other pre-existing proficiency targets that may be salient to parents and teachers. In this appendix, we first provide evidence in support of our multiple-target formulation, before assessing alternative modeling choices that all involve a single target.

Table B.1 provides motivating evidence, giving 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 suggests that schools did respond to more than the middle target. Specifically, 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 B.1 – Fractions of students scoring above low, middle (proficiency), and high (superior performance) targets, by grade and year

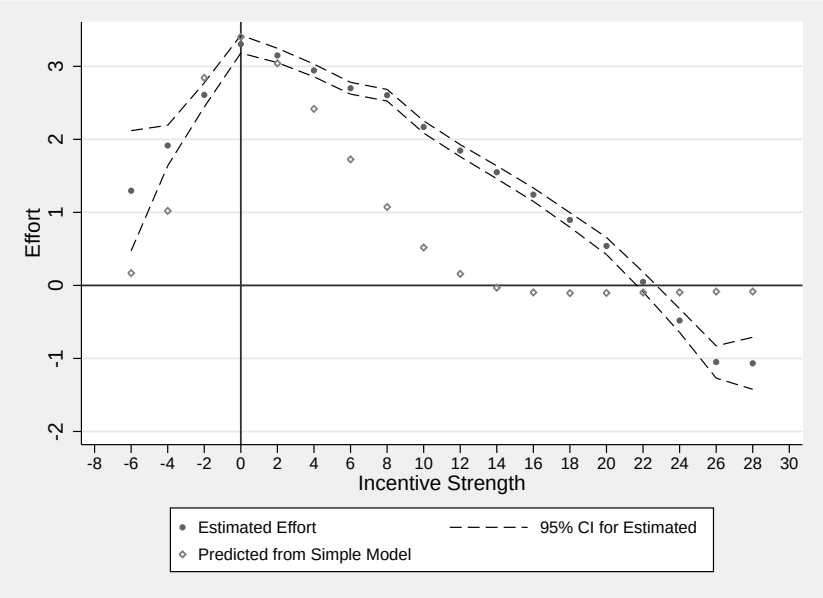
	Year		
Grade and Target	2001-02	2002-03	2003-04
Third Grade			
Low	96.9%	98.9%	98.9%
Proficiency	78.3%	89.2%	89.2%
High	37.1%	44.8%	45.1%
Fourth Grade			
Low	99.1%	99.3%	99.2%
Proficiency	89.5%	94.9%	94.6%
High	45.8%	60.5%	60.6%
Fifth Grade			
Low	98.3%	98.9%	99.1%
Proficiency	89.1%	92.7%	93.5%
High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

As shown in Figure 5 (in Section VII.B), our multi-target-response model produces a very good fit 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 will indicate this is not the case, based on the implied fit alongside the estimated effort function.

B.1. A Benchmark: Single-Target Scheme

As a benchmark, consider the simple single-target model, shown in Figure B.1. It is apparent that the fit is poor, particularly in terms of the effort devoted to high-achieving students.

Given the single-target model, 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 noted when discussing the comparative static properties of optimal effort. Yet this would be at odds with the value of σ^2 that is identified outside the teacher's problem, depending 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 1, along with the 95 percent confidence intervals, and the binned means of effort implied by the single-target scheme.

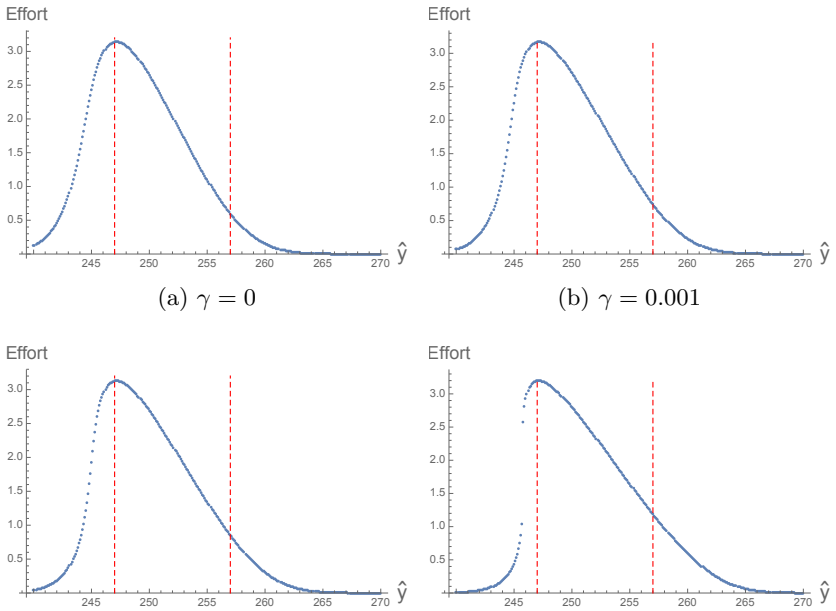
FIGURE B.1 – INVERTED-U RESPONSE TO NCLB AND MODEL FIT OF SINGLE-TARGET SCHEME

In contrast, as Figure 5 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. It is worth considering whether alternative modeling choices could achieve the same end under a single-target scheme, as we assess next.

B.2. Complementarity in Production

One potential alternative 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 the coefficient on the interaction term (γ) influences the shape of the implied effort profile. Figure B.2 simulates the



(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 5).

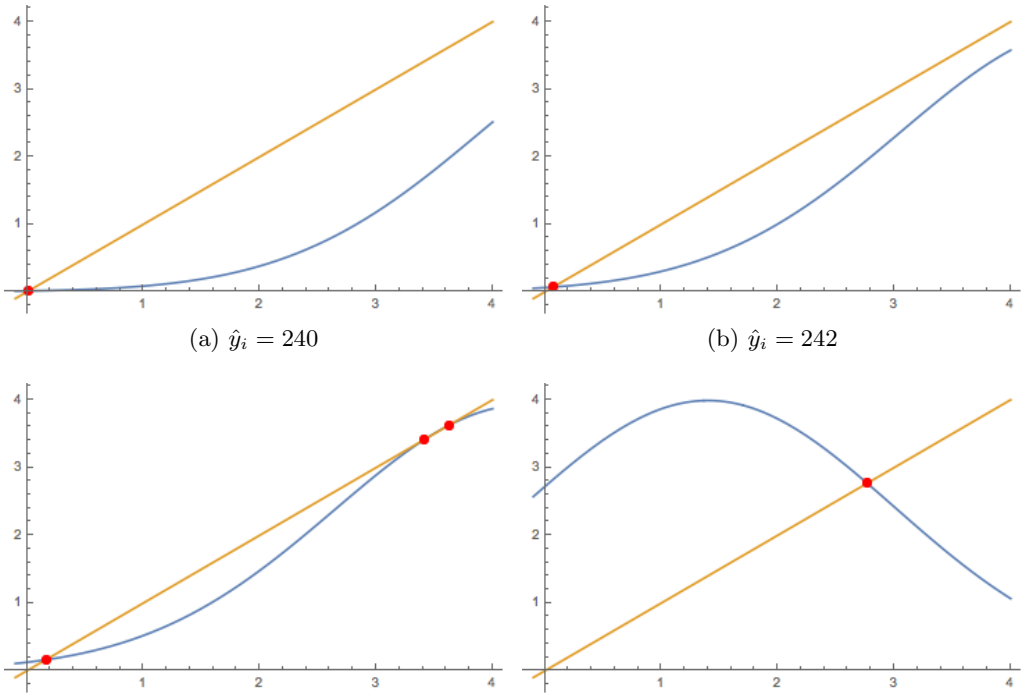
FIGURE B.2 – SIMULATION OF EFFORT FUNCTION FOR DIFFERENT DEGREES OF COMPLEMENTARITY

effort profile for different values of γ.[69] Relative to the case in which there is no complementarity (Figure B.2(a), as in our actual model), increasing γ does broaden the right-hand side of the effort profile. Yet doing so would require a very large value of γ, which in turn would ensure that the two profiles would not match on the left (see the bottom two panels, where effort declines precipitously to the left of the first vertical dashed line, representing the NCLB target y_M^T).

The second consideration relates to the way the marginal benefit and cost curves intersect for different student types (based on 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 values of $\hat{y}_i$ above a threshold level (given the properties of the effort-setting model).[70] Such jumps start to occur in our model for values of $\gamma = 0.0046$, or greater. 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.

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

[70] Figures B.3(a) and B.3(b) demonstrate this feature clearly for $\gamma = 0.006$, showing that a low-ability student with $\hat{y}_i = 240$ or $\hat{y}_i = 242$ receives close-to-zero effort, while a low-ability student with a fractionally higher predicted score, $\hat{y}_i = 243$ or above, receives substantially higher effort.



(a) $\hat{y}_i = 240$ (b) $\hat{y}_i = 242$

(c) $\hat{y}_i = 243$ (d) $\hat{y}_i = 250$

FIGURE B.3 – SIMULATION OF OPTIMAL EFFORT ($MB = MC$) FOR $\gamma = 0.006$

B.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.

B.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^2_ν. 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 their students quite accurately, leading us to rule out prediction error as a viable alternative explanation.

C. Properties of the Optimal Effort Solution

In this appendix, we 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.C). 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.[71] 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 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 (6) 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,[72] 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 level of incentive strength. In words, two different feasible parameter vectors cannot both yield the same global maximum in terms of the teacher objective across all student 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

[71] 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.

[72] We compute the sum of effort for the average teachers classroom by calculating the implied effort for each student in a classroom (according to her $\hat{y}$ and the effort profile) and then summing those values across all students in the classroom.

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 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 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.[73] 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.[74] 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.[75] 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.

[73] 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.

[74] 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.

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

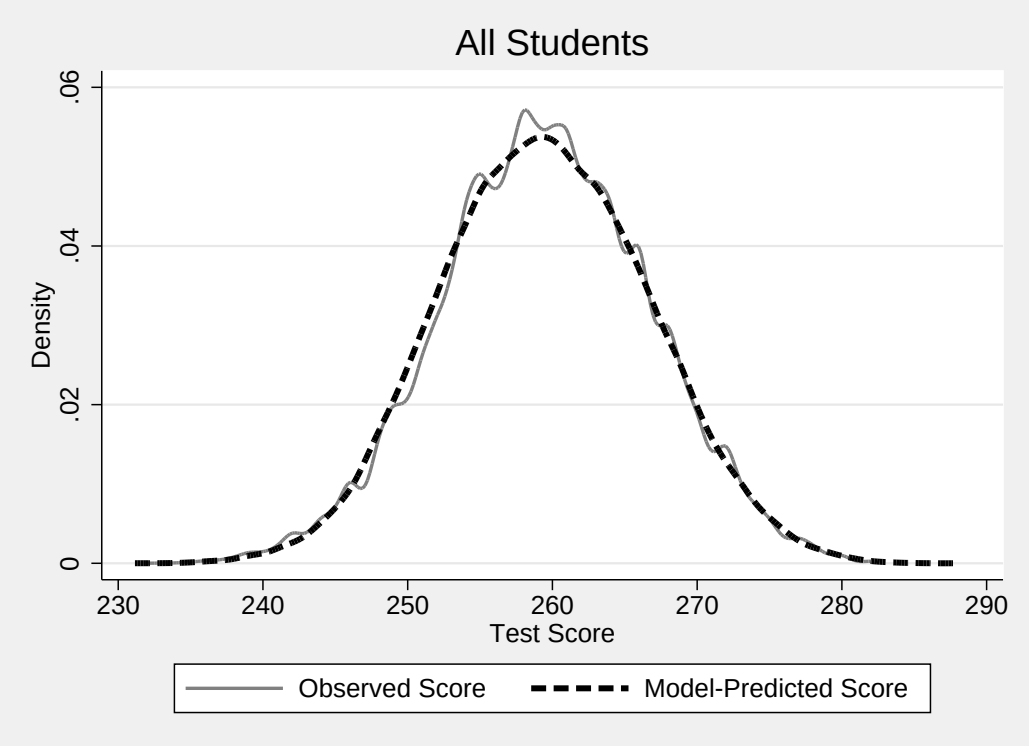
D. Computation Appendix

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 (5) 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.C, this results in N_c first-order conditions for each classroom, given by equation (6), 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 (6) indeed maximizes the teacher's objective.[76] 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.

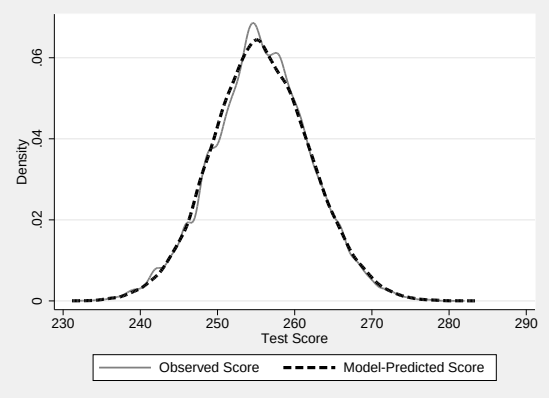
[76] Because *fminunc* is a minimization routine, we apply it to the negative of the teacher's objective function, ensuring the recovered solution maximizes the objective function.

E. Model Fit

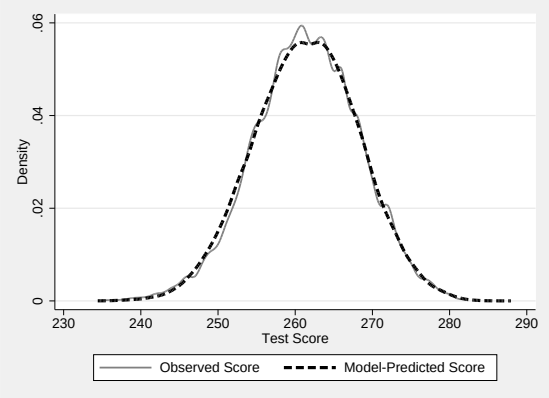


Notes: This figure presents the density of observed test scores (measured in developmental scale units) and the density of test scores predicted by the model for the full sample.

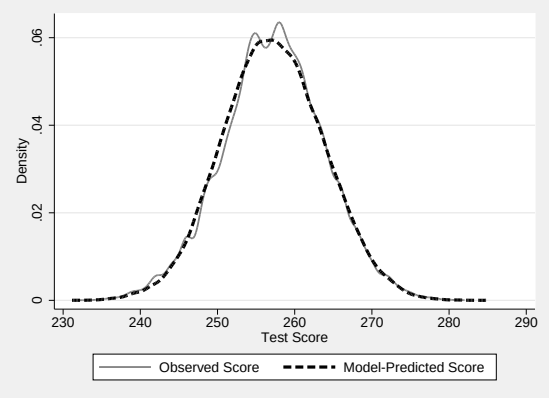
Figure E.1 – Distributions of Observed and Model-Predicted Scores



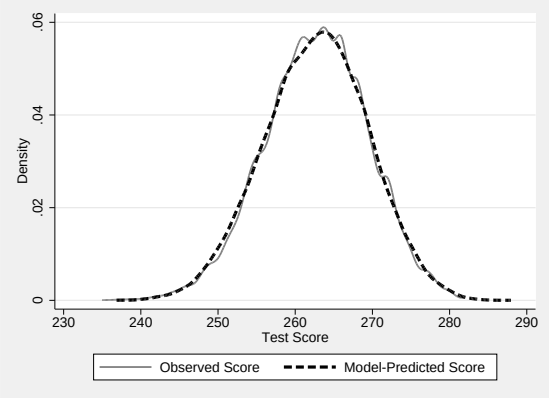
(a) Black Students



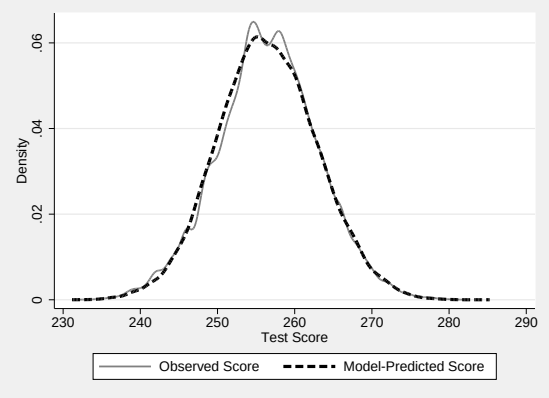
(b) White Students



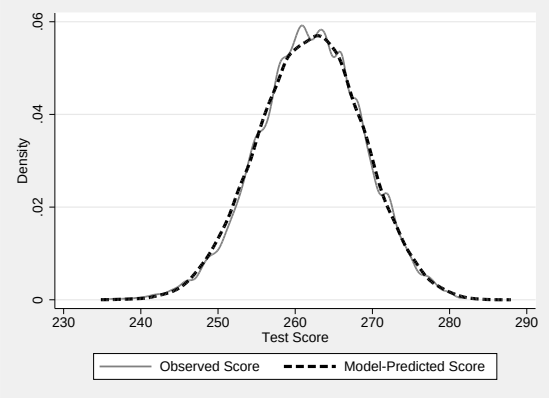
(c) Not College-Educated Parents



(d) College-Educated Parents



(e) Economically Disadvantaged



(f) Not Economically Disadvantaged

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

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

F. Simulation Appendix

This appendix provides background to the counterfactual simulations. In it, we describe 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 under NCLB, 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 the 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 $\hat{y}$).[77]

To illustrate the form of these two heterogeneous bonus payment parameterizations, Figure F.1 shows each of them as a function of predicted scores along with the baseline homogeneous bonus payment case in which $w_i = 1, \ \forall \ i$, 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. 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.[78] 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 $y_{it}^T = \delta + \alpha y_{it-1}$, 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 $(\hat{y}_{it} - y_{it}^T)$ 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.

Along those lines, we take each fixed target in the set Y^f in turn in our simulations. 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 the student VA targets matches the value (in developmental scale points) of a given fixed target in the set Y^f. For example, suppose we are matching the real NCLB fixed target of 247. In that case, we set $\delta = 247 - \alpha\bar{y}_{t-1}$, which implies that the mean VA target is also 247.

[77] Here, the bonus payment b_M is scaled by the student-specific weight $w_i = w^H(\hat{y}_i) = \frac{(\hat{y}_i-y_{\min}+1)}{\hat{y}_{\text{med}}-\hat{y}_{\min}+1}$.

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

We conduct this exercise for each fixed target in Y^f, looping through successive values of $y^T \in Y^f$. 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.[79]

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

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

where N_t is the total number of students in the state, $\widehat{\beta} \equiv [\widehat{\frac{b_M}{\psi}}, \widehat{d}_M, \widehat{\theta}, \widehat{\sigma}^2]$,[80] and $\epsilon_i \sim N(0, \widehat{\sigma}^2)$. Note that the set R can be either drawn from the family of fixed targets, in which case each student has the same target, $y_{it}^R = y^R$, $\forall\ i$, and y^R is an element of the set Y^{fixed} above, or from the family of VA targets, in which case each student has a student-specific target given by $y_{it}^R = \delta^R + \alpha^R y_{it-1}$.

To explain the cost-equating procedure, 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 ψ 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 $b(k)$ 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 $y^T = 247$.[81]

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 (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

[79] To see this, note that we have $y_{it}^T = \delta + \alpha y_{it-1} = y^T - \alpha\bar{y}_{t-1} + \alpha y_{it-1}$, thus implying that the mean value of y_{it}^T (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 $\bar{\hat{y}}_t - y^T$.

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

[81] 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.

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

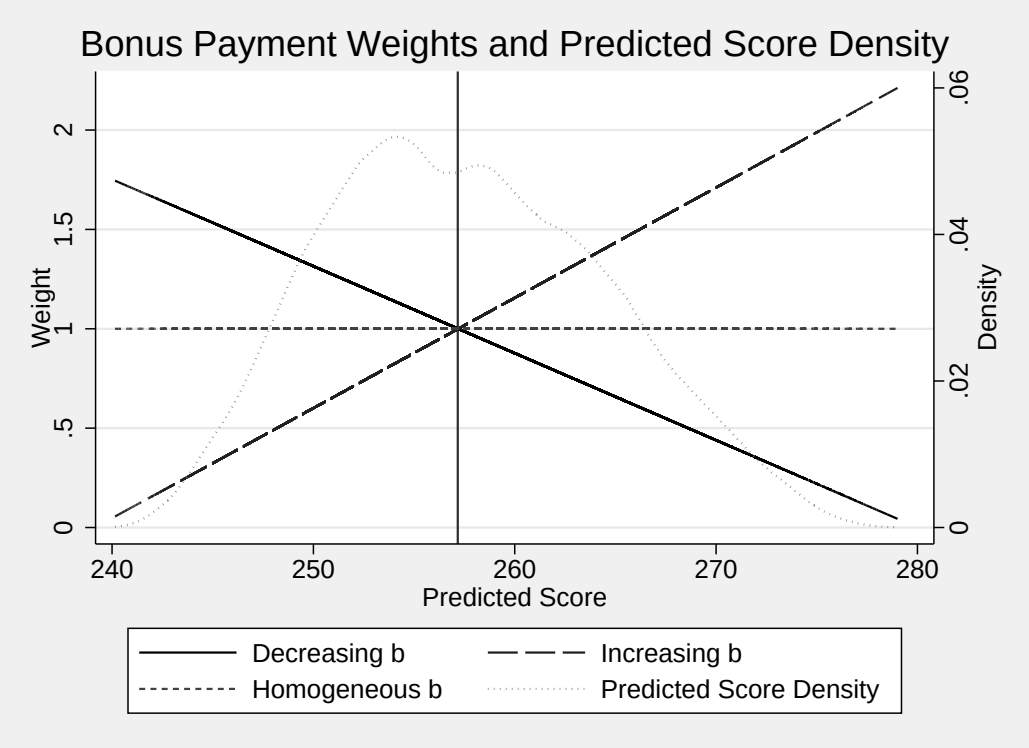
Under heterogeneous bonus payment regimes, average costs under the scheme that places more weight on low-performing students and the scheme that places more weight on high-performing students are determined by

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

and

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

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[82] 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 6.

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.[83] 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

[82] 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.)

[83] 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

Target (Percentile Position)	(1) Mean Effort DE	(2) Mean Effort CEE	(3) Inverse of the Test Score Variance DE	(4) Inverse of the Test Score Variance 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 6 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. 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.[84] 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, simply changing the bonus payment regime generates the outward shift in the frontier, causing both an increase in mean effort and a reduction in test score variance.

For higher targets on the frontier, cost-equating across regimes is required to increases mean effort when switching bonus payment regimes. At relatively high proficiency targets,[85] the new bonus regime creates tension between the incentive to devote effort to (relatively) high-performing students and the incentive to devote effort to low-performing students due to the heterogeneous bonus payments. As a result, 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.[86] Lower (unadjusted) mean effort under the new regime implies that costs are too low. Thus, in order to equate

[84] For example, at 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).

[85] More specifically, simply switching regimes results in mean effort increases for targets up to the 20th percentile of the predicted score distribution, after which point cost-equating is needed to increase mean effort and generate the frontier's outward shift.

[86] 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) after switching regimes, 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.

costs with the NCLB benchmark, a higher bonus payment must be offered to increase teacher effort and the proficiency rate. In these cases, the change to aggregate effort following from cost-equating more than offsets the loss from the new bonus structure, increasing mean effort above the original value under heterogeneous bonus payments and resulting in an outward shift in the frontier.

Table G.3 precisely quantifies these forces that cause the shift out of the frontier in Figure 7. In the table, we distinguish between the effects that arises purely from switching to the new bonus payment structure without ensuring cost equivalence – i.e., the 'bonus payment effect' – and the effect that then arises subsequently when equating costs – i.e., the 'cost-equating effect.'[87] For each target we consider, the magnitude of each effect is calculated relative to the prevailing mean effort and inverse test score variance under that same target in the homogenous bonus payments case.

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 Effect	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 bonus payment effect and the cost-equating effect on mean effort, respectively, that occur in Figure 7 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 magnitude of each effect on the inverse of the test score variance.

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

As Figure 7 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, both proficiency target incentives and bonus payment incentives are strongest for students who are high

[87] 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.

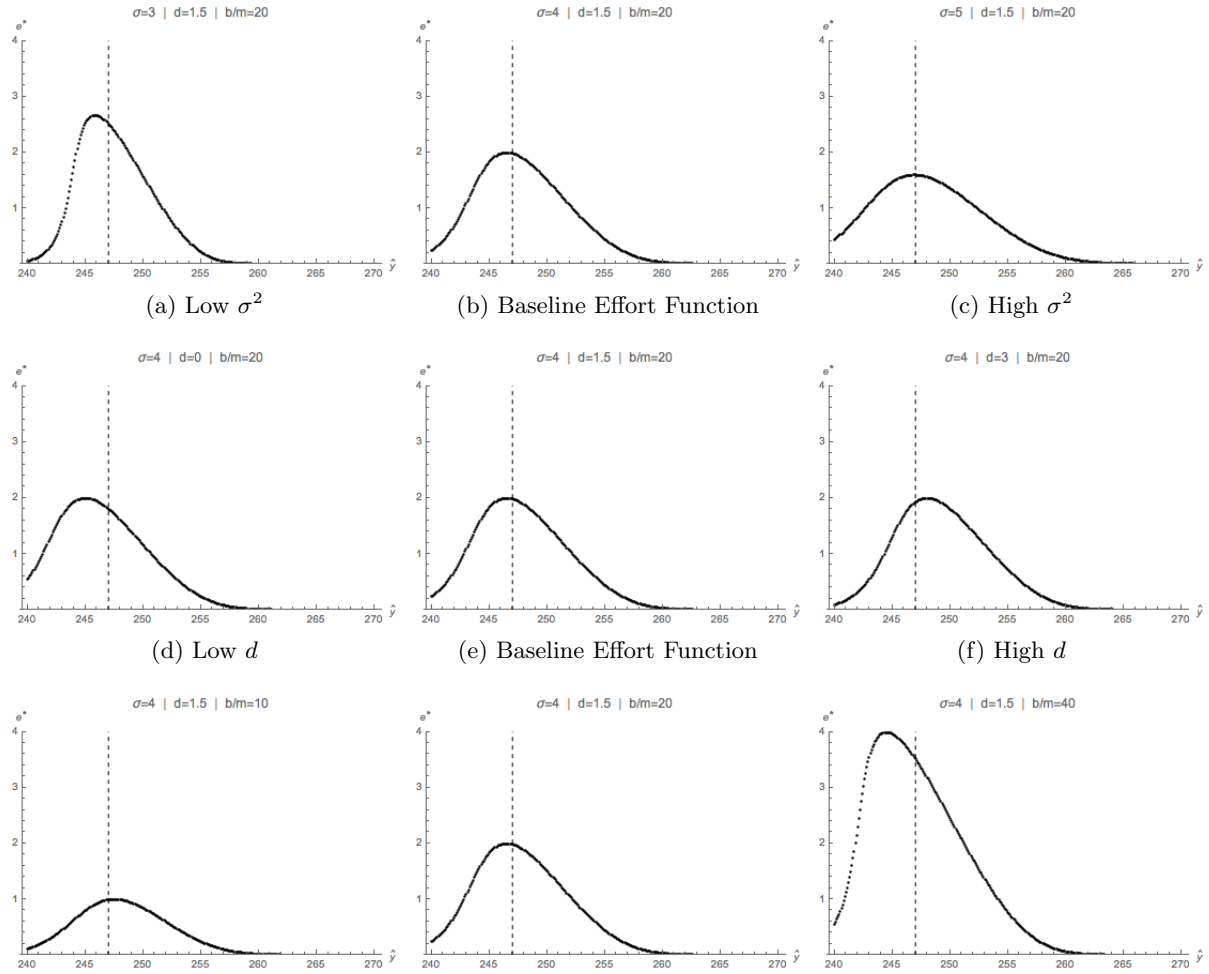
in the predicted test score distribution. These students receive the largest amount of extra effort, while low-performing students experience the largest reduction. Because the strongest students experience test score gains and weakest experience losses, inequality (test score variance) also rises. Together, 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 d (e) Baseline Effort Function (f) High d

(g) Low $\frac{b}{\psi}$ (h) Baseline Effort Function (i) High $\frac{b}{\psi}$

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, focusing on a single target (the NCLB target, indicated by the vertical dashed line). 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 the 'shift' parameter d rises; panels (g) through (i) show how the *height* of the profile increases as the scaled benefit of passing, $\frac{b}{\psi}$, rises.