NBER WORKING PAPER SERIES

INCENTIVE DESIGN IN EDUCATION: AN EMPIRICAL ANALYSIS

Hugh Macartney
Robert McMillan
Uros Petronijevic

Working Paper 21835
http://www.nber.org/papers/w21835

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2015

We would like to thank Joseph Altonji, Peter Arcidiacono, David Deming, Giacomo De Giorgi, David Figlio, Caroline Hoxby, Lisa Kahn, Lance Lochner, Rich Romano, Eduardo Souza-Rodrigues, Aloysius Siow, and seminar participants at Duke University, University of Florida, NBER, SITE, University of Western Ontario, and Yale University for helpful comments and suggestions. Thanks also to Hammad Shaikh for excellent research assistance. Financial support from the University of Toronto is gratefully acknowledged. All remaining errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Incentive Design in Education: An Empirical Analysis
Hugh Macartney, Robert McMillan, and Uros Petronijevic
NBER Working Paper No. 21835
December 2015
JEL No. D82,I21,J33,M52

ABSTRACT

While incentive schemes to elicit greater effort in organizations are widespread, the incentive strength-effort mapping is difficult to ascertain in practice, hindering incentive design. We propose a new semi-parametric method for uncovering this relationship in an education context, using exogenous incentive variation and rich administrative data. The estimated effort response forms the basis of a counterfactual approach tracing the effects of various accountability systems on the full distribution of scores. We show higher average performance comes with greater score dispersion for a given accountability scheme, and that incentive designs not yet enacted can improve performance further, relevant to education reform.

Hugh Macartney
Department of Economics
Duke University
239 Social Sciences Building
Box 90097
Durham, NC 27708
and NBER
hugh.macartney@duke.edu

Robert McMillan
University of Toronto
Department of Economics
150 St. George Street
Toronto, ON M5S 3G7
CANADA
and NBER
mcmillan@chass.utoronto.ca

Uros Petronijevic
University of Toronto
150 St. George Street
Toronto, Ontario, Canada
M5S3G7
uros.petronijevic@utoronto.ca

Across many types of organization, schemes that provide incentives to exert effort are often seen as an important means of boosting organizational performance. The design of such schemes has, naturally, been a central preoccupation in economics and also a very challenging one, given that effort is typically unobserved. While this challenge has been taken up in a substantial body of sophisticated theoretical research,[1] a host of incentive schemes operating in practice are only loosely informed by the theoretical contracting literature. This creates scope – in a variety of settings – for potentially significant performance gains from judicious incentive reform.

In order to gauge whether such gains are attainable, one attractive approach involves studying the introduction of actual incentive reforms, as in classic papers by Lazear (2000) and Bandiera, Barankay and Rasul (2005), for example.[2] Incentive designers often wonder about more speculative considerations, however, looking to the effects of changing the parameters of existing schemes counterfactually, or the effects of incentive schemes yet to be implemented in practice. As a complement to the evaluation of actual schemes, therefore, approaches that combine a strategy for identifying effort under prevailing incentive provisions with a framework for counterfactual analysis can be appealing – a type of approach that features in a recent body of research studying worker incentives.[3]

We build on that strand of literature in two key respects. First, we propose a new semi-parametric method for recovering the incentive strength-effort relationship. Here, we specify a simple model of effort setting, then show that the implied function relating effort to incentives can be credibly and transparently identified, based on exogenous incentive variation; despite its simplicity, the model does a remarkably good job of fitting the data. Second, using the estimated model as a foundation, we develop a framework for carrying out informative counterfactuals, allowing us to uncover the full distribution of outcomes for

[1] The seminal work of Mirrlees (1975) and subsequent analyses make clear that robust contract forms are difficult to obtain.

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

[3] See innovative papers by Copeland and Monnet (2009) and Misra and Nair (2011), among others.

a given incentive scheme and further, to place rival incentive schemes on a common footing; here, we include a maximally efficient benchmark – a scheme that generates the highest average effort for a given cost.[4] The framework thus allows us to explore how changing incentives counterfactually affects the complete distribution of outcomes.

We develop our new approach in the context of accountability systems in public education – a prominent policy arena in which incentive schemes have been adopted widely. In their essence, such schemes involve setting performance targets and explicit rewards (or penalties) that depend on target attainment, their goal being to increase teacher and school effort, in turn raising test scores.[5] Several types of accountability scheme have been implemented to date, including proficiency schemes – notably the federal No Child Left Behind Act of 2001 ('NCLB') – that set fixed performance targets based on school sociodemographics, and value-added schemes whose targets condition on prior student scores.

Such variety brings to mind important incentive design issues, particularly how the features of different accountability systems affect student and school outcomes. While several convincing studies consider incentive issues by focusing on particular aspects of accountability schemes already in operation,[6] our approach allows us to analyze the impacts of alternative education accountability systems, including ones not yet implemented, across the entire distribution of test scores. In so doing, we are able to assess – for the first time – the relative merits of rival schemes in a quantifiable way, and shed light on the way incentives can affect the spread of educational outcomes, relevant to wider social mobility.

To implement our approach, we exploit plausibly exogenous incentive variation arising from the introduction of NCLB. Being a 'fixed' scheme, NCLB creates incentives to focus on students at the margin of passing relative to a fixed target.[7] We take advantage of this non-uniformity in North Carolina, a setting for which we have rich administrative data covering all public school students over a number of years.

[4]Our approach has a positive emphasis, in contrast to the normative emphasis of the optimal contracting literature.

[5]Persuasive evidence that accountability schemes succeed in improving student achievement already exists – see Carnoy and Loeb (2002), Lavy (2009), Hanushek and Raymond (2005), Dee and Jacob (2011), and Imberman and Lovenheim (2015), among others.

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

[7]That NCLB creates such incentives has been well-documented in the literature – see Reback (2008), for example.

To guide our empirical analysis, we set out a simple model of the education process that links incentives to outcomes via discretionary action, 'effort,' which we will take to refer to changes in observable test scores that are attributable to incentive variation. This yields an effort function that depends on the parameters of the incentive scheme and, under threshold targets, a measure of incentive strength.

We estimate this function semi-parametrically, first by constructing a continuous incentive strength measure for each student using rich data from the pre-NCLB-reform period, equal to the gap between the target and the student's predicted score; this describes how marginal each student is.[8] Then we compare the achievement of each student against a prediction reflecting all pre-reform inputs; this difference for each level of incentive strength serves as a pre-period performance control. Once incentives are altered, teachers and schools re-optimize, and the post-reform difference between the realized and predicted test scores will reflect both the original inputs as well as any additional effort associated with the new non-uniform incentives, likely to be strongest where students are marginal.[9]

Consistent with the model predictions, we find that the profile of actual scores in the pre-reform period plotted against the incentive measure is remarkably flat. Then, once the reform comes in, there is a pronounced hump, peaking precisely where incentives should be most intense and declining on either side of that.[10] By differencing the post- and pre-reform distributions, we can then uncover – based on minimal assumptions[11] – the underlying effort response to greater accountability for all levels of incentive strength.

This response forms the basis of a counterfactual approach that allows us to recover the outcome distribution under various accountability schemes, including schemes not currently

[8]Given the incentive strength measure is very much related to, and builds upon, measures appearing in related prior work, we draw attention in Appendix A to seemingly subtle differences that turn out to be important in the development of our approach.

[9]This is related to the approach in Neal and Schanzenbach (2010). Because they only have one year of pre-reform data, they do not construct a score difference relative to the pre-reform; and unlike their focus on deciles of the distribution, we develop a *continuous* measure of incentive strength. (See Appendix A for more discussion.)

[10]We also find evidence supporting the hypothesized channel, rather than the rival story of schools focusing on the middle of the distribution.

[11]We assume first that the education production technology is linear in effort and separable – a reasonable first-order approximation, made almost without exception in the education literature. Second, NCLB should influence the effort decisions of educators but not the other determinants of student test scores – something we can check indirectly.

implemented. To illustrate the approach, we trace out comparable performance frontiers based on the first two moments of the score distribution for different accountability systems, starting with the most widespread – those setting fixed and value-added targets.[12] Doing so reveals a clear tradeoff: higher average performance comes at the expense of more dispersed scores for a given type of scheme. Further, we show that the frontier associated with fixed targeting falls inside its value-added counterpart, and that regular value-added schemes are inferior to a scheme with student-specific bonus payments. We are able to represent these respective frontiers together with the aid of a single diagram. The analysis is relevant to the reform of existing education accountability systems, and has broader applicability to the problem of incentive design in the workplace – a theme we develop below.

The rest of the paper is organized as follows: the next section sets out a theoretical framework that underlies our estimation approach. In Section II, we describe the institutional context and the rich administrative data set we have access to, along with motivating descriptive evidence. Section III presents the research design: we outline our implementation of this design in a North Carolina context in Section IV, along with estimates of the effort function. In Section V, we describe our counterfactual framework, and present the results from the counterfactual analysis. Section VI concludes.[13]

I. Theory

We present a simple model of the education process that links accountability incentives to school performance. This motivates our incentive strength measure, and serves as a means for analyzing the determinants of optimal effort, which will guide our empirical implementation. We also consider the setting of accountability targets – an important issue in incentive design.

The model has three main elements: First, there is a *test score technology* relating measured education output y to various inputs. Given our interest in incentives, we place particular emphasis on the discretionary actions of educators that may serve to increase output. In line with a substantial body of work in incentive theory, we will refer to such

[12] Given we predict the full counterfactual score distribution, our approach allows us to consider various other possibilities beyond the first two moments.

[13] Supplementary material is provided in a set of appendices, referenced in the text.

actions simply as 'effort.'[14] In our setting, effort is taken to capture a range of actions on the part of educators that raise student performance, most of which are unobserved by the researcher.[15]

Formally, we write the education production technology, which relates outputs to inputs, as $y_i = q(e_i, \theta_i) + \epsilon_i$. We focus on the effort choice of a single educator i; depending on the context, this could be a single teacher, or all the teachers in a grade within a school, treated as a single effort-making body. Output y_i is the test score of the student or students taught by i; e_i is the effort of educator i, which is endogenous to the prevailing incentive scheme; and θ_i represents exogenous inputs, such as student ability – we treat students as passive, though potentially heterogeneous. Effort and exogenous student inputs are assumed to be related in a systematic way to output, captured by the function $q(e, \theta)$, with both first partial derivatives being positive; thus we will think of increases in θ – relevant below – as capturing more favorable exogenous 'production' conditions. The simplest form for the $q(.,.)$ function would be linear, and additive in its arguments. The output measure y_i is also assumed to be influenced by a noise component, given by ϵ_i. We define $H(\cdot)$ and $h(\cdot)$ as the cumulative distribution and probability density functions of the negative of this noise, $-\epsilon_i$, and assume these functions are common across all educators i.

Second, we characterize an *incentive scheme* by a target y_i^T faced by educator i and a reward b, both exogenously given.[16] This formulation of the target allows for a range of possibilities, considered in more detail below: the target could be an exogenously fixed score, a function of average student characteristics (including past performance), or even be student-specific. The reward parameter b governs how target attainment maps into the educator's payoff, and can include monetary rewards or non-monetary punishments.

Third, educator i faces a cost that is convex in effort and may depend on exogenous

[14]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."

[15]In our empirical implementation, effort will refer more specifically to changes in observable test scores that are attributable to incentive variation.

[16]Unlike the contract theory literature, we will not write down a planning problem and then derive the optimal contract form given the incentive constraints. Rather, we focus on the way that the effort distribution changes under exogenously given schemes, where those schemes might not be optimal.

conditions θ also, though we suppress that dependence for now. Thus we write it as $c(e_i)$, and assume its functional form is known.[17]

Taking these elements together, we can write down the educator objective under different incentive schemes. In each instance, the structure allows us to express optimal effort as a function of the parameters of the incentive scheme, along with other exogenous characteristics. This function is our main object of interest, our goal being to recover its form semi-parametrically: in Section III, we outline a strategy for doing so.

Types of Scheme

We consider some common incentive schemes through the lens of the model. Under a *piece rate*, educator i's objective can be written: $U_i = b[q(e_i, \theta_i) + \epsilon_i - y_i^T] - c(e_i)$. Optimal effort e_i^* then satisfies the first-order condition: $b \cdot \frac{\partial q(e_i, \theta_i)}{\partial e_i} = c'(e_i)$, implying that the choice of effort is invariant to the target y_i^T.

In practice, *threshold* schemes are far more widespread in an education setting, not least because they give policymakers greater cost control – the maximum payout under the scheme is determinate, for example. We focus on these. The educator's objective under a threshold-based scheme is $U_i = b \cdot 1_{y_i \geq y_i^T} - c(e_i)$, which in expectation is given by $b \cdot \Pr[q(e_i, \theta_i) - y_i^T \geq -\epsilon_i] - c(e_i) = b \cdot H[q(e_i, \theta_i) - y_i^T] - c(e_i)$. Optimal effort e_i^* will then implicitly satisfy the first-order condition, given by

$$b \cdot h[q(e_i, \theta_i) - y_i^T] \frac{\partial q(e_i, \theta_i)}{\partial e_i} = c'(e_i). \tag{1}$$

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

Take the case where $q(e, \theta)$ is simply the sum of its arguments: $q(e_i, \theta_i) = e_i + \theta_i$. The additively separable assumption implies that the marginal benefit of effort, given by the LHS of (1), simplifies to $b \cdot h[\theta_i + e_i - y_i^T]$. Holding the reward parameter b fixed, marginal benefit is then a function of two quantities. The first is the gap between the systematic component

[17] This formulation applies straightforwardly in the case of a single educator teaching a single student. When multiple students are involved, it is natural to distinguish two components of effort, one being student-specific, the other common to all students taught by i. Below, we will consider the polar cases of student-specific versus common effort.

of the score, $q(e_i, \theta_i)$ and the target for educator i; the second is the density of the error in the performance measure, $h(\cdot)$, evaluated at that gap.

For illustration, suppose that the decision maker i refers to a school. Suppose also that the error term is unimodal, peaking at a mean of zero. Then consider three cases, corresponding to variation in the underlying conditions governing education production in three types of school, where $\theta_L < \theta_M < \theta_H$ – for instance, those educating low-, moderate- and high-ability students on average, respectively.

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

Given our interest in the agents' decision problem, consider the school in the first case, where $\theta = \theta_L$. Taking the target as fixed, the low value of θ makes it very unlikely that the school will exceed its performance target, even if effort is set at a high level; thus, the incentive to exert costly effort will be correspondingly low. We illustrate this case in panel (a). The marginal benefit curve, conditioning on θ_L, will simply be the product of (fixed) b and the density $h(\cdot)$, tracing out the shape of the latter. Optimal effort, $e^*(\theta_L)$, is determined by the intersection of this marginal benefit curve and the given marginal cost curve.

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

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

Several relevant points emerge from this simple illustration: First, a given target can create stronger or weaker incentives, depending on the underlying 'production' conditions facing the educator. Intuitively, a target will engender more effort when effort has a higher marginal impact in terms of the passing probability – where the density of the noise distribution is higher.

Second, there is a clear role for incentive design to strengthen effort incentives: if θ is given exogenously and is known, then it would be possible to tailor targets to create the strongest possible incentives (given θ) if all that mattered were maximal performance.[18] Related, we will see shortly that different methods of setting targets will give rise to different incentives – our empirical exploration of the associated incentive differences will form the heart of the paper.

Third, a natural metric for measuring incentive strength emerges from the analysis. This is the gap between the systematic component of the score and the target – the argument of the $h(\cdot)$ function in the expression for the marginal benefit of effort. It will feature in our empirical implementation with one important adjustment: we will replace the systematic component of the score, $q(e, \theta)$, with a predicted score, $\hat{y}$, obtained using parameters estimated in a low-stakes environment.

Fourth, in line with the illustration, the optimal effort response is likely to have an inverted-U shape.[19] For low and high values of θ, incentives to exert effort will be low, as increased effort has little impact on target attainment: in the former case, the target is likely to remain out of reach, while in the latter, it is easily attained. In the middle of the θ distribution, effort incentives are stronger. We will see below that the spread of the effort distribution will be influenced by the way targets are chosen, with implications for incentive

[18] In panel (b) of Figure A.1, the target calls forth precisely the *maximal* effort, given that the marginal cost curve intersects the marginal benefit curve at its highest point.

[19] In the simple model under consideration, having a symmetric, unimodal error distribution is sufficient for this.

design.

Target setting

Having discussed the simple analytics of target threshold schemes in general, we now consider different target-setting schemes. A general class of incentive schemes can be characterized by a set of targets and rewards, written $\{y^T(I), b(I)\}$, given prior information I. Relevant to our application, we will treat the predicted score using all prior information available to the econometrician ($\hat{y}$) as our summary of I. Thus the general characterization can be expressed as $\{y^T(\hat{y}), b(\hat{y})\}$. The schemes most widely used in practice, discussed next, can be viewed as special cases.

Fixed schemes: These involve targets that are the same for all agents – say, those involved with a certain grade g. Let the test score of students taught by educator i in grade g be given by y_{ig}.[20] The grade-specific target y_{ig}^T that applies under a fixed scheme can be written y_g^T. The fixed scheme sets a threshold: $b \cdot 1_{y_{ig} \geq y_g^T}$, where b is the reward if test score y_{ig} exceeds the student-invariant target y_g^T, or the sanction if the score does not exceed the target (e.g. under NCLB).

Value-added schemes: The target now depends on prior information. The threshold rule can be written $b \cdot 1_{y_{ig} \geq y_{ig}^T}$ where the target $y_{ig}^T = \alpha_g y_{ig-1}, \ \forall \ i$ in grade g. The parameter α_g is central to the target-setting process in grade g, governing the strength of the dependence on the prior score; b is the reward if the test score y_{ig} exceeds the target, or the sanction if the score does not exceed the target.

To illustrate the general formulation, suppose for simplicity that performance in the prior grade, y_{g-1}, is the only information available, and let the *predicted* score be written as a flexible function $\hat{y} = \hat{\alpha}_0 + \sum_{p=1}^{P} \hat{\alpha}_p (y_{g-1})^p$, dropping any person-specific subscripts, where the parameters $\{\hat{\alpha}_p\}_{p=0}^{P}$ are estimated from a flexible regression of y on y_{g-1} in a low-stakes incentive environment.[21]

In this setup, the target under a fixed scheme imposes $\alpha_0 = \alpha$, and $\alpha_p = 0$ for all $p > 0$.

[20] For simplicity, assume one such student.

[21] As in schemes typically implemented, the reward $b(y_{g-1}) = b$ for all prior information y_{g-1}. Let the target be calculated according to $y^T(y_{g-1}) = \alpha_0 + \sum_{p=1}^{P} \alpha_p (y_{g-1})^p$ for some $\{\alpha_p\}_{p=0}^{P}$.

Correspondingly, the target under a value-added scheme imposes the condition $\alpha_p = 0$ for $p > 1$.

Uniform schemes: What we will refer to as a *uniform* scheme also serves as a useful reference point. Given the definitions, y^T can replicate $\hat{y}$ with some constant shift d. In particular, if $\alpha_0 = \hat{\alpha}_0 - d$ and $\alpha_p = \hat{\alpha}_p$ for all $p > 0$, then $y^T = \hat{y} - d$. The *maximally efficient* scheme, which provides an important benchmark in our counterfactual analysis, is a special case of this. (See Section V.)

I.A. Optimal Effort Function

Our main object of interest from the model is the optimal effort function, which solves the educator's payoff maximization problem. For a given incentive scheme, involving a target y_{ig}^T set in a specific way and bonus b, this can be written as $e^*(\hat{y}_{ig} - y_{ig}^T; b)$ for an educator i in grade g. We already saw, in the general case of threshold schemes, how the gap between the systematic component of the score and the target is a potentially important determinant of the effort decision. Thus, we write optimal effort as a function of the gap: $\hat{y}_{ig} - y_{ig}^T$.[22]

Because this gap plays a key role in the empirical implementation that follows, it is useful to define our continuous measure of *incentive strength* as $\pi_{ig} \equiv \hat{y}_{ig} - y_{ig}^T$ for a given type of scheme. The distribution of incentive strength will also be important: under a value-added scheme, it is likely to be tighter than under a fixed scheme, given that targets can be made student-specific.

With this compact notation in place, our interest centers on $e^*(\pi_{ig})$ – the way that a given measure of incentive strength, observed for educator i in grade g, maps into effort, taking b as given. It is worth emphasizing that the functional form of $e^*(\pi_{ig})$ is *unknown* and must be inferred empirically: Section III describes our strategy for uncovering this mapping in a semi-parametric way.

[22]From equation (1), it is clear that the reward parameter b enters the marginal benefit expression multiplicatively. We will suppress b for the remainder of the section, though this multiplicative feature will be useful when carrying out counterfactual analysis.

Aggregation: The Case of Uniform Effort

Here, we provide a brief discussion of aggregation issues, which will arise later on in the empirical analysis. The simple version of the model presented thus far applies most readily to the case of a single student taught by single effort-making educator. This can be thought of as capturing the extreme case where the teacher is able to perfectly tailor instruction to each student. In such a setting, it is easy to aggregate up to the classroom or school level, useful when exploring the classroom or school-wide effects of incentive schemes.

This extreme case does not do justice to the constraint that teachers often face, whereby it is difficult to individualize instruction. As an alternative, we consider the other extreme classroom aggregation case where each teacher chooses an identical level of effort for all students under her care: Under the additive separability assumption, it is straightforward to show that the educator chooses a level of uniform effort according to the distance between *average* student ability and the target averaged over all students in the class.[23] This formulation of the educator's decision problem, when she is constrained to choose a common level of effort for all of her students, will prove useful when we present our empirical results.

II. Institutional Setting and Data

North Carolina provides a suitable context for our study, for institutional and data reasons. On the institutional side, the state provides incentive variation arising under two separate accountability regimes. High-stakes accountability was implemented under the ABCs of Public Education legislation in the 1996-97 school year for all schools serving kindergarten through grade eight. Under the ABCs, each grade from three to eight in every school is assigned a school-grade-specific target gain, depending on both average prior student performance and a constant level of expected test score growth. Based on average school-level gains across all grades in student standardized mathematics and reading scores, the ABCs pays a monetary bonus to all teachers and the principal if a school achieves its overall growth target.

NCLB provisions were implemented in North Carolina in the 2002-03 school year fol-

[23]It is also worth noting that class/school size effects do not play a role in the uniform effort choice under additive separability.

lowing the passage of the federal No Child Left Behind Act in 2001. In contrast to the pre-existing pecuniary incentives under the ABCs, NCLB focuses on penalties for under-performing schools. The federal program aims to close performance gaps by requiring schools to meet Adequate Yearly Progress ('AYP') targets for all students and for each of nine student subgroups. We focus on AYP targets for all students as a first approximation to the prevailing incentives, abstracting from the subgroup aspect in our analysis.

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

Our sample period runs from 1997-2005. To focus on schools facing similar incentives, we limit the sample to schools serving kindergarten to eighth grade, and exclude vocational, special education, and alternative schools.[24] These restrictions notwithstanding, our sample sizes are very large, with over five million student-grade-year observations over the nine-year window, and over 14,000 school-year observations.

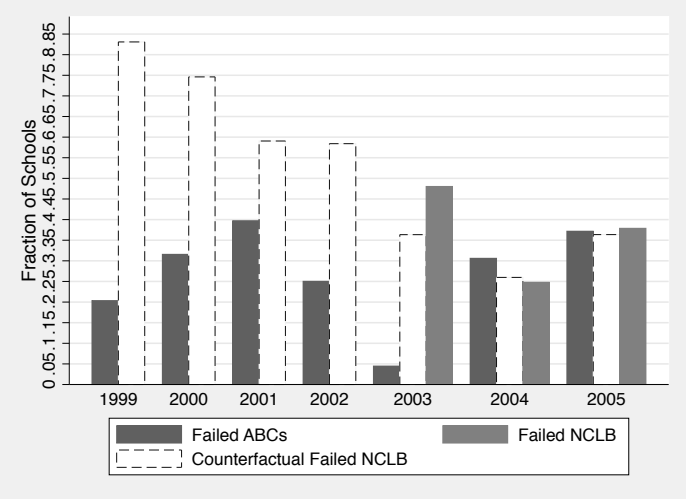
Table A.1 provides sample summary statistics. Our main performance variables are constructed from individual student test scores. These are measured on a developmental scale, which is designed so that each additional point represents the same amount of knowledge gained, irrespective of the baseline score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

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

With respect to the school-level performance variables, 27 percent of schools failed the ABCs and 37 percent failed NCLB across the sample period.[25] Throughout our sample period, the average school had a school-wide proficiency rate of 79 percent on math and reading tests.[26]

II.A. Descriptive Analysis

We are interested in school and school-grade performance variation over time, especially contrasting outcomes before and after the introduction of NCLB. In prior work, in order to identify the effects of NCLB on student outcomes, Dee and Jacob (2011) use states that had pre-existing accountability programs as 'control' states, and argue quite plausibly that NCLB should have had little effect on school incentives there. To motivate our research design below, we present evidence indicating that schools in North Carolina – a state with a pre-existing scheme – actually responded in a noticeable way to the introduction of NCLB.



Notes: The figure shows (a) the fraction of schools failing the ABCs (all years), (b) the fraction of schools failing NCLB (2003-), and (c) the fraction of schools *predicted* on a consistent basis to fail NCLB, from calculations using the underlying student-level data (all years).

FIGURE 1 – SCHOOL PERFORMANCE FROM 1999 TO 2005

Figure 1 shows the fraction of schools failing the ABCs and NCLB in each year, starting in 1999. We construct a consistent series showing the counterfactual NCLB failure rate in

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

[26]The school-wide proficiency rate does not map directly into school-level NCLB outcomes because of subgroup proficiency requirements.

the years prior to 2003 by applying the NCLB outcome calculations in 2003 to the underlying student-level performance data for prior years. As is clear from the figure, the year in which NCLB was introduced – 2003 – is associated with a remarkable decline in the fraction of schools failing the ABCs, down from 25 percent in 2002 to only 4 percent in 2003. To lend credence to the notion that this reflects an NCLB-triggered response, notice that the fraction of schools predicted to fail NCLB declines substantially, from 58 percent in 2002 to 36 percent in 2003, consistent with schools taking steps to improve along the dimensions required under NCLB.[27] The figure also shows that the ABCs improvement was short-lived, as the failure rate more than jumps back, to over 30 percent in 2004.

III. Research Design

The research design presented in this section is central to our analysis. Our goal is to uncover the optimal effort response $e^*(\pi)$ for a given incentive strength π, described in Section I. Building on the descriptive evidence in the previous section, the strategy we follow makes use of the new performance requirements under NCLB as an exogenous shock to the school decision process occurring in 2003. In order to explain the semi-parametric approach we develop, we set out the technological assumptions we are making, describe the construction of our *ex ante* incentive strength measure, and then show how double-differencing combined with an exogeneity argument yields the optimal effort response to incentives.

Technology: As is standard in the literature, we specify a simple linear structure for the test score production function. Not only does this provide a useful starting point; it also serves as a reasonable first-order approximation to a richer underlying test score technology.

We think of there being a 'pre-reform' environment in which effort is approximately uniform, irrespective of incentive strength π.[28] Test scores in this environment are generated according to $y(\pi) = \hat{y} + \epsilon(\pi)$, the sum of a systematic component, which may include baseline effort, and noise. We reference a particular score by our *ex ante* incentive measure

[27] Due to the many nuances associated with the implementation of NCLB, we are unable to perfectly reproduce school-level outcomes in the post-NCLB period: the counterfactual NCLB failure rate does not coincide precisely with the actual failure rate. In 2003, we understate the failure rate by around 10 percentage points: in 2004 and 2005, we are much closer.

[28] Such uniformity can be checked, to some degree. We provide descriptive evidence in the next section (see Figure 6).

π, as we are interested in seeing how changes in formal incentives are reflected in the score distribution in a way that is attributable to an effort response.

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

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

expressed as a function of π.

Incentive Strength and Effort Response: We obtain the optimal effort response as a function of incentive strength using a five-step procedure (visualized in Figure A.3), in which we distinguish 2003 – the year in which the new incentives came into effect – from pre-reform years.

In the first step (Figure A.3a), we predict student performance in a flexible way in those pre-reform years using several covariates, including lagged test scores.[29] In the second step (Figure A.3b), we then use the saved coefficients from the first step to construct our *ex ante* incentive strength measure. In particular, combining those coefficients with updated covariates from 2003 and prior test scores for 2002, we are able to predict performance ($\hat{y}$) for 2003. Using the known NCLB target specified by the reform (y^T), we then compute our continuous measure of incentive strength as the difference between the predicted value (which does not include *additional* effort in 2003) and the target: $\pi \equiv \hat{y} - y^T$. On this basis, the predicted score component is invariant to any changes occurring in 2003. Instead, variation in incentive strength when new incentives are considered arises from changes in the target. Specifically, the proficiency target y^T becomes relevant under NCLB, implying that π will capture the strength of effort incentives in 2003 but not in prior years.

With the *ex ante* incentive strength measure in hand, we then turn to the main task

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

in the third step: to determine the optimal effort response for each value of our continuous incentive strength measure (π). We do this in the following semi-parametric way: for each value of π, we compute the difference between the realized and predicted scores ($y - \hat{y}$) in 2003, the year when the incentive shock occurred. Intuitively, this quantity should contain the effort response as well as any noise in our prediction.[30] In terms of the above technology, this step will recover $y_R(\pi) - \hat{y}_R = e^*(\pi) + \epsilon_R(\pi)$. Recall from the theory that the response to NCLB is predicted to be non-uniform (larger for marginal students with π close to zero and smaller for non-marginal students with larger absolute values of π). This prediction serves as a helpful check that our assumptions are satisfied.

Our fourth step is designed to control for any pre-existing patterns that are common across the pre- and post-reform periods. To that end, we repeat steps one through three using only the pre-reform years 1998 through 2000 (Figures A.3c and A.3d).[31] That is, we regress 1999 scores on cubics in prior 1998 math and reading scores as well as contemporaneous student covariates. Using the resulting coefficients, we construct $\hat{y}$ and, combined with the target y^T, π for the year 2000. We then compute $y - \hat{y}$ in 2000 for each value of the incentive measure. This fourth step thus recovers the noise in the pre-reform period ($y(\pi) - \hat{y} = \epsilon(\pi)$).

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

$$\big(y_R(\pi) - \hat{y}_R\big) - \big(y(\pi) - \hat{y}\big) = e^*(\pi) + \epsilon_R(\pi) - \epsilon(\pi)\,. \tag{3}$$

In our context, an exogeneity assumption implies that the RHS of (3) is just equal to $e^*(\pi)$, the desired object. That is, conditional on π, the stochastic components of the production technology are equal in expectation over time. Given that NCLB should influence the effort decisions of educators but not the other determinants of student test scores, this assumption is plausible – recall that the targets under NCLB are student-invariant.[32] We consider

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

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

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

supportive evidence next.

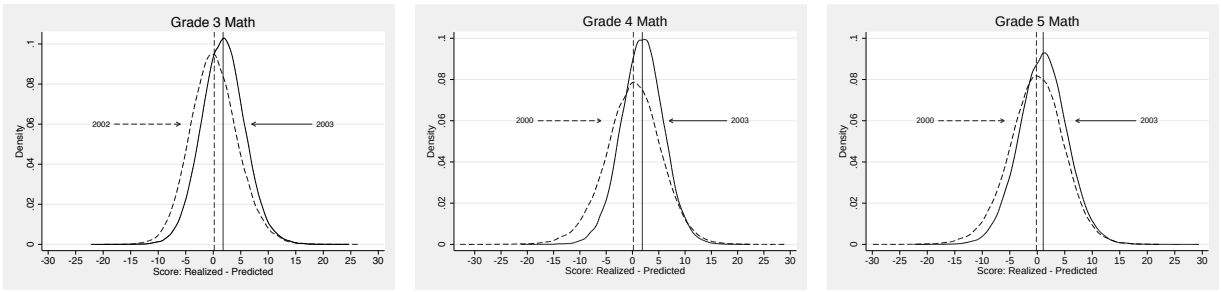
IV. Incentive Response

In this section, we show results from the implementation of our research design, and discuss evidence relating to the validity of the approach.

IV.A. The Test Score Response

Given our rich test score microdata, we can compute whether there was any test score response to the introduction of NCLB in 2003.

Figure 2 shows the densities of realized minus predicted test scores in both the pre-period (2000 and 2002)[33] and the post-period (2003), which we interpret as the densities of unobservable test score determinants, including the effort of educators. Predicted scores



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

Figure 2 – Shifts in Residual Densities (2003 versus 2000 and 2002)

represent the test scores that are likely to occur in a given year if the relationship between student observable characteristics and realized test scores remains the same as it was in past years. The difference between realized and predicted scores in the pre-NCLB year is centered approximately around zero, suggesting that the prediction algorithm performs well. In 2003, however, the residual densities for all grades display clear rightward shifts, indicating that

[33]For grades four and five, we use 2000 as the pre-reform year, rather than the year immediately preceding the implementation of NCLB (2002). We do so because North Carolina altered the scale used to measure end-of-grade results in 2001, implying that we cannot use our prediction algorithm in 2002, as contemporaneous and prior scores are on different scales in 2001. In contrast, we can use it in 2002 for grade three, because these students write the 'pre' test at the beginning of the year, meaning that both scores are on the same scale in 2001.

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

Interpreting the Response as Effort

We argue that the unobserved determinant in question is teacher effort by relying on the well-established theoretical predictions associated with proficiency-based accountability schemes, discussed in Section I. These schemes reward schools (or refrain from punishing them) based on the percentage of proficient students and so provide schools with clear incentives to focus their efforts on students predicted to score around the proficiency target. Students likely to score far below the target require a prohibitively costly amount of extra effort to reach proficiency status, while students predicted to score far above the target are likely to pass without any additional effort at all. Thus, to the extent that the documented shifts in residual densities represent an effort response, we should see the largest gains in realized-over-predicted scores for the students predicted to score near the proficiency threshold.

Figure 3 shows that these are exactly the patterns we find across the predicted test score distribution.[34] In 2003, the gains above predicted scores are low for students predicted

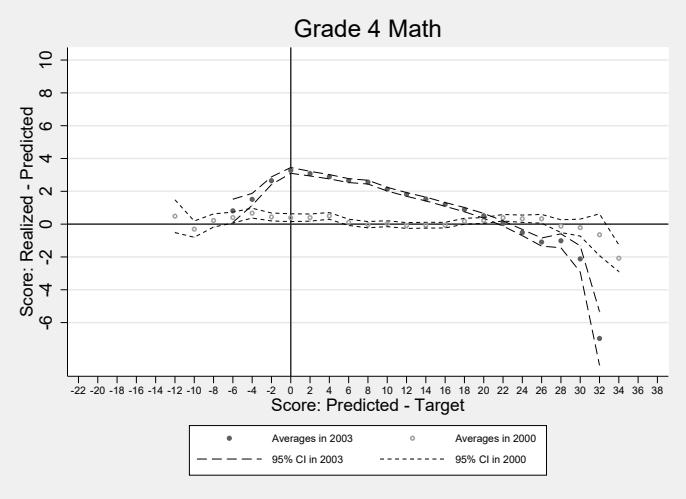


FIGURE 3 – NCLB EFFORT RESPONSE

[34]We focus on the fourth grade distribution in our analysis. The patterns are similar across grades for the middle and top of the ex ante incentive strength distribution, though the low end is somewhat distorted for third and fifth grades due to slightly greater heterogeneity in the impacts of the pre-existing value-added ABCs scheme as well as effects that can be traced to subgroups under NCLB.

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

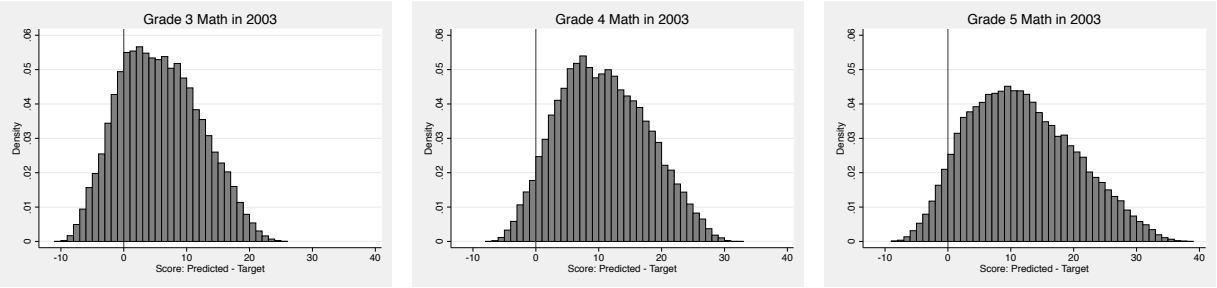
Generally speaking, to the extent that educators care about incentives under the new regime, adjusting discretionary effort is an obvious candidate input through which performance can be altered, and in a manner consistent with the observed change in the test score profile. This 'effort' could take a variety of unobserved forms: teachers raising their energy levels and delivering material more efficiently inside the classroom, increasing their lesson preparation outside the classroom, or teaching more intensively 'to the test.' Without richer data, these various components are difficult to distinguish. At the school level, there could be changes in education spending (in the form of lowering class sizes, for instance), or reassigning teachers, though we do not find evidence of observable changes along these dimensions. This leads us to take the evidence as supporting the view that teachers changed their effort in response to NCLB, and in a way according with the hypothesized response to a proficiency-count system.

IV.B. Validity of the Approach

Testing for Bunching

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

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



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

FIGURE 4 – DISTRIBUTION OF PREDICTED SCORES MINUS THE NCLB TARGET

In each of the panels, the fixed NCLB target occurs at zero, as indicated by the vertical line. Based on the distribution of predicted scores, the figure provides no evidence of bunching. This lends support to the notion that the NCLB 'shock' was indeed exogenous, affecting the effort of educators but not other determinants of student test scores.

Rival Effort Story

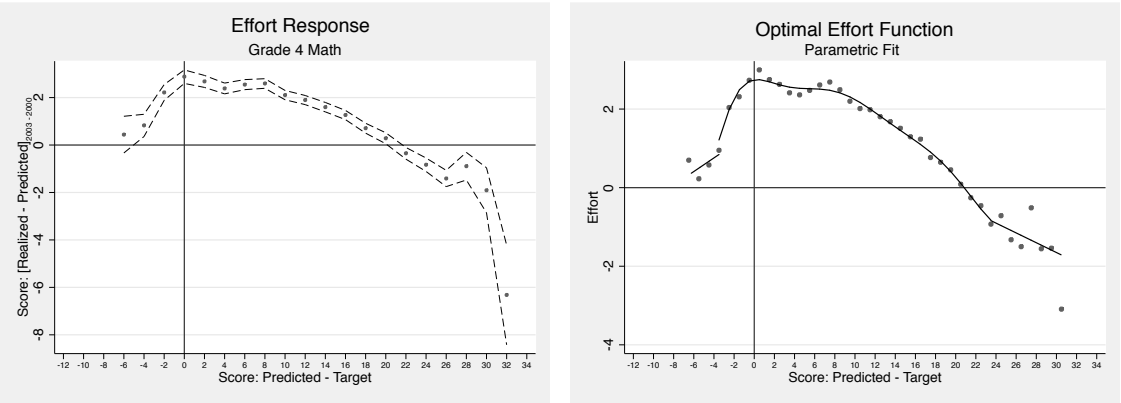
Our maintained hypothesis is that our approach uncovers the effort response to the incentive strength measure, π. In Appendix B, we consider alternative story whereby effort might vary with respect to a student's relative position in the predicted score distribution in his or her school. A natural test is to look at the position of peak effort across quartiles of the incentive strength distribution, as we describe there, supporting the view that schools indeed respond to a student's proximity to the proficiency threshold, as hypothesized.

IV.C. Incentive Strength and the Optimal Effort Function

We are now in a position to recover the effort function. In Figure 5(a), we take the difference between the two years – 2003 versus 2000 – to isolate the effort response at each point in the predicted test score distribution (applying the third and fourth steps). In Figure 5(b), we then fit a flexible polynomial to the data, which we interpret as the optimal effort function, $e^*(\pi)$. We estimate the function by first grouping students in each year into incentive strength bins of width one (in terms of developmental scale units) and calculating the average effort response within each bin. We then take the difference between the 2003

and 2000 averages, weight each difference by the number of students in the bin, and regress the weighted differences on a flexible tenth-order polynomial of the incentive strength measure, π. The points in Figure 5(b) represent the within-bin differences and the function is the tenth-order polynomial fit. To avoid over-fitting noisy outcomes in bins with relatively few observations, we impose a linear fit on the effort function in the extreme positive and negative ranges of π.

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



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

FIGURE 5 – DERIVATION OF THE OPTIMAL EFFORT FUNCTION

Rationalizing the Estimated Effort Function

We show how the estimated effort function can be rationalized using a simple parametrization of the basic model of optimal teacher effort-setting in Section I.

Writing teacher preferences as $U_i = b \cdot 1_{y_i \geq y^T} - c(e_i, \theta_i)$, we allow educator i's cost of effort to depend here on student characteristics, θ_i, proxied using the predicted score $\hat{y}_i$; thus we write the systematic portion of the production technology $q(e_i, \theta_i) = q(e_i, \hat{y}_i)$. We also assume $q(e_i, \hat{y}_i) = e_i + \hat{y}_i$, and $c(e_i, \hat{y}_i) = \frac{1}{2} s(\hat{y}_i) e_i^2$, further parametrizing the slope of the marginal cost of effort as $s(\hat{y}_i) = \psi_0 + \psi_1 \hat{y}_i + \psi_2 \hat{y}_i{}^2$.

The first-order condition in equation (1) can then be used to solve for optimal effort

implied under this parametrization, estimating the parameters of the marginal cost function using a minimum distance estimator. Figure A.2 shows the simulated optimal effort against the continuous incentive strength measure, $\pi = \hat{y} - y^T$, on the horizontal axis alongside the estimated effort profile recovered using our semi-parametric approach.

The figure makes clear that the simple quadratic parametrization of the marginal cost function in the basic model can reproduce the estimated effort function remarkably well, including its right-skewed shape. This suggests that the model provides a surprisingly good approximation to the effort-setting process, despite its abstracting from many features of the actual incentive environment.

V. Counterfactual Analysis

In this section, we first present our framework for carrying out counterfactuals, before turning to the counterfactual results themselves.

V.A. A Framework for Counterfactual Comparisons

We develop a framework that enables us to compute the implied test score distribution for the set of targets $\{y_{ig}^T\}$ associated with a given incentive scheme, whether actual or counterfactual.[35] This implied score distribution serves as a basis for calculating useful summary measures – the score distribution's first and second moments, for example. These measures are then used to trace out performance frontiers associated with a given type of scheme by varying the scheme's targets in a counterfactual way, and also to compare frontiers across various types of schemes (e.g., fixed versus value-added). We will do so with the aid of a simple figure that allows us to represent different classes of scheme on the same picture in an informative way.

The basic component of the framework is a test score technology of the form given in equation (2). Consider a counterfactual scheme R and its associated target (or targets). To determine the test score for educator i under the scheme R, we define the test score outcome

[35] Recall that we characterize a scheme in terms of a set of targets and rewards.

as the sum of the predicted test score, the educator's effort,[36] and a noise component:

$$y_i(R) = \hat{y}_i + e^*\big(\pi_i(R)\big) + \epsilon_i. \tag{4}$$

In terms of the elements of this formula, we already have estimates of effort as a function of incentive strength $e^*(\pi)$, and the noise ϵ is drawn from a mean-zero normal distribution with variance matching the year 2000 distribution in panel (b) of Figure 2. This leaves the implied distribution of incentive strength measures for scheme R, $\pi_(R)$. We describe how this is recovered next.

Distribution of Incentive Strength: Under target-based schemes, for a given distribution of exogenous attributes faced by educators and a given target, we can calculate the distribution of distances, one for each educator i, between the predicted score and the target. Computationally, this requires calculating $\pi_{ig} = \hat{y}_{ig} - y_{ig}^T$ for educator i in grade g, and storing the entire set of distances across all educators.

Taking the two components of incentive strength in turn, a scheme R will imply a determinate target for each educator. This can be compared to the educator's *predicted* score, which we assume – looking ahead to our actual implementation – is calculated in the 'pre-reform' environment referred to above. Thus, while changing the incentive scheme will change the individual target and alter the optimal effort calculation (as well as the actual score, if effort also changed), it will not alter the predicted score used in the calculation of the distribution of incentive strength.[37]

By way of illustration, Figure 4 showed grade-specific density plots of the incentive strength measure implied by the prevailing NCLB targets for each grade. As the target is moved counterfactually, above and below the actual target, the density will shift to the left and to the right, respectively.[38]

Formally, we will write the density of π under scheme R as $f_R(\pi)$ on the support

[36] Below, we will consider two cases, where educator effort is student-specific or classroom-specific, starting with the former.

[37] The fact that the predicted score is defined not to include any effort response will be reflected in the notation used in this subsection.

[38] Focusing on fourth grade, these movements for a fixed scheme are shown in panels (a) and (e), relative to (c), in Figure A.5.

$\pi \in [\underline{\pi}_R, \bar{\pi}_R]$. The incentive strength calculations for each educator (just outlined) allow us to recover this density semi-parametrically.

Counterfactual Outputs: Our framework yields the full counterfactual score distribution for any scheme R.[39] This allows us to compute a variety of useful summary measures. For illustration, we focus on the average effort associated with a given scheme and the corresponding variance of scores. These two measures can then be used to trace out performance frontiers for each type of scheme, which we then plot.

Given the definition of the incentive strength measure from above (and omitting subscripts), we write *average effort* for a given incentive scheme R as $\Omega_R \equiv \int_{\underline{\pi}_R}^{\bar{\pi}_R} e^*(\pi, b) f_R(\pi) d\pi$.

For comparability, we will use the average cost of a scheme to standardize outcomes across schemes. The total cost under a given scheme is the monetary reward associated with a student passing, multiplied by the number of such students, assuming that this is the relevant payoff structure. To compute the average cost, first define $\tilde{\pi}_{ig} \equiv \pi_{ig} + \epsilon_{ig} = y_{ig} - y_{ig}^T$; this is the gap between the target and the *actual* score, given that π_{ig} does not include the noise component. Then the average cost for scheme R can be written $C_R = b \int_{\underline{\tilde{\pi}}_R}^{\bar{\tilde{\pi}}_R} \mathbf{1}_{(\tilde{\pi}+e^*(\pi,b)\geq 0)} \tilde{f}_R(\tilde{\pi}) d\tilde{\pi}$, where $\tilde{f}_R(\tilde{\pi})$ is the distribution of $\tilde{\pi}$ given scheme R, with the lower and upper limits given in the integral.[40] This is just the proportion of educators predicted to exceed the relevant target under R multiplied by the reward parameter b.

This formula allows us to compute one of our preferred summary measures, *average effort* under scheme R with targets $\{y_R^T\}$ *relative to* scheme R' with targets $\{y_{R'}^T\}$, which can be calculated using Ω_R, $\Omega_{R'}$, C_R and $C_{R'}$, just defined. One might wish, as we will, to compare average effort under fixed versus the maximally efficient benchmark (defined shortly) – i.e. $\Omega_{R=F}$ versus $\Omega_{R'=M}$. This cannot be accomplished directly, however, if the average costs under the two schemes differ (i.e., $C_M \neq C_F$). Our solution is to adjust b under the fixed scheme until $C_F(b') = C_M(b)$, and then compare $\Omega_M(b)$ to $\Omega_F(b')$. We will use the measure, $\frac{\Omega_M(b)}{\Omega_F(b'),}$ to compare performance under these alternative schemes.

[39] Specifically, for a scheme R and for each educator i, we can now compute the implied test score $y_i(R)$ according to (4), once we have determined the incentive strength $\pi_i(R)$ faced by that educator under the scheme, and read off the corresponding effort level from the estimated effort function. Repeating for all educators under scheme R yields the full counterfactual outcome distribution.

[40] Note that e^* is added to $\tilde{\pi}$ since $\hat{y}$ does not include effort under regime R.

Now turning to the dispersion of scores under scheme R, define $\tilde{y} \equiv \hat{y} + \epsilon$. In essence, this is the test score absent any effort component, referring back to (2). The variance in test scores is then:

$$\Sigma_R = \int_{\tilde{y}_R}^{\bar{\tilde{y}}_R} [\tilde{y} + e^*(\hat{y} - y^T, b) - \overline{\tilde{y} + e^*}]^2 \tilde{f}_R(\tilde{y}) d\tilde{y},$$

where we make explicit the dependence of e^* on $\hat{y}$. The ratio $\frac{\Sigma_F(b)}{\Sigma_{VA}(b')}$ can be thought of as *relative dispersion*, or the change in variance from adopting the less sophisticated target.[41]

Alternative Types of Scheme: For the set of *fixed* proficiency targets, we simply move the proficiency threshold below and above the NCLB threshold. For *value-added* ('VA') proficiency targets, we vary the multiplicative coefficient on the prior mathematics scores in the ABCs math target, moving it below and above the coefficient that actually prevails under the ABCs.[42]

Under *uniform schemes*, first referenced in Section I, the target $y^T = \hat{y} - d$, where d is some constant shift. This means that the incentive strength measure for a uniform scheme will simply be $\pi = d, \ \forall \ \hat{y}$. Average effort is then $e^*(d)$, since $e = e^*(d), \ \forall \ \hat{y}$, and average cost is $C(d) = b \int \mathbf{1}_{d+e^*(d)+\epsilon \geq 0} h_\epsilon(\epsilon) d\epsilon$.[43] This allows us to define the *maximally efficient* scheme by choosing d to maximize $\frac{e^*(d)}{C(d)}$.

Although uncommon in practice, the reward b can also be distributed *heterogeneously* for any given target. This is a further possibility that we explore in our counterfactual analysis. Generally, a heterogeneous reward has implications for both average effort and the dispersion of scores, but under a uniform scheme it only affects the dispersion. Here, increasing the weight on students with the lowest $\hat{y}$ will serve to lower the spread. If $e^{max} \ll \sigma^2_{\hat{y}}$, then there will be a positive lower bound for the variance, which we will plot below.[44]

[41] Intuitively, greater effort is delivered under the value-added scheme to students who were considered non-marginal under the fixed scheme, making effort more uniform across the student distribution under the former.

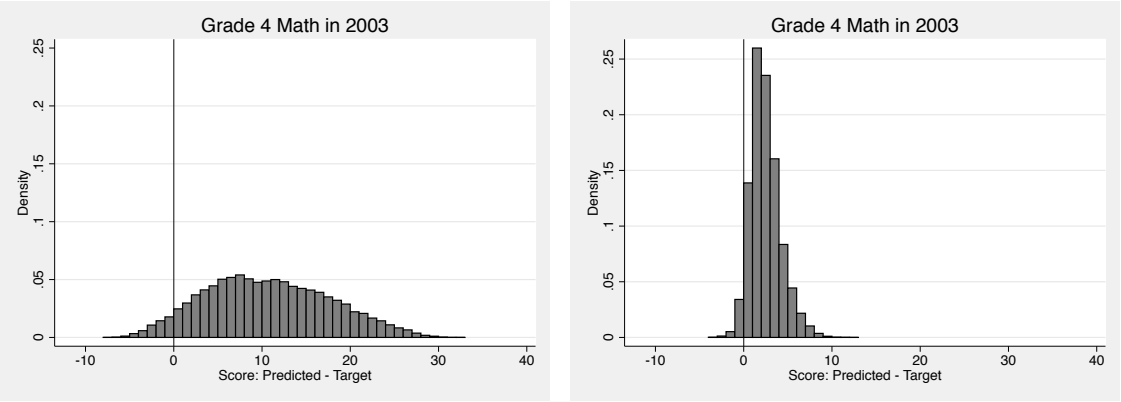
[42] When we consider determining proficiency status based on a VA target, we assume all of the other rules under NCLB still hold but that, instead of students facing a common proficiency threshold, each student faces a individual-specific proficiency threshold equal to his or her VA target determined by the ABCs formula, with the multiplicative adjustment just described.

[43] In Section I, we define $h(\cdot)$ as the probability density function of $-\epsilon$. Here, we use $h_\epsilon(\cdot)$ to denote the probability density function of ϵ.

[44] Using a heterogeneous reward to reach that minimum variance under the maximally efficient scheme is socially optimal only if it is sufficiently transparent to agents; otherwise, the effort response may be

V.B. Counterfactual Results

In describing our counterfactual results, we first consider a set of fixed proficiency targets and a set of VA proficiency targets, deriving the implied incentive strength distributions for π in each case. To illustrate the proficiency targets that we consider, Figure 6 shows the distributions of π that prevail under the actual NCLB proficiency target and the ABCs VA target. The figure underlines the point that since VA targets are student-specific, the distribution of π under the ABCs target has a much lower variance than the distribution under the NCLB target. An important implication of this is that the effort responses under VA targets will be much more uniform across the distribution of students than those under fixed targets.



(a) NCLB Target (b) ABCs Target

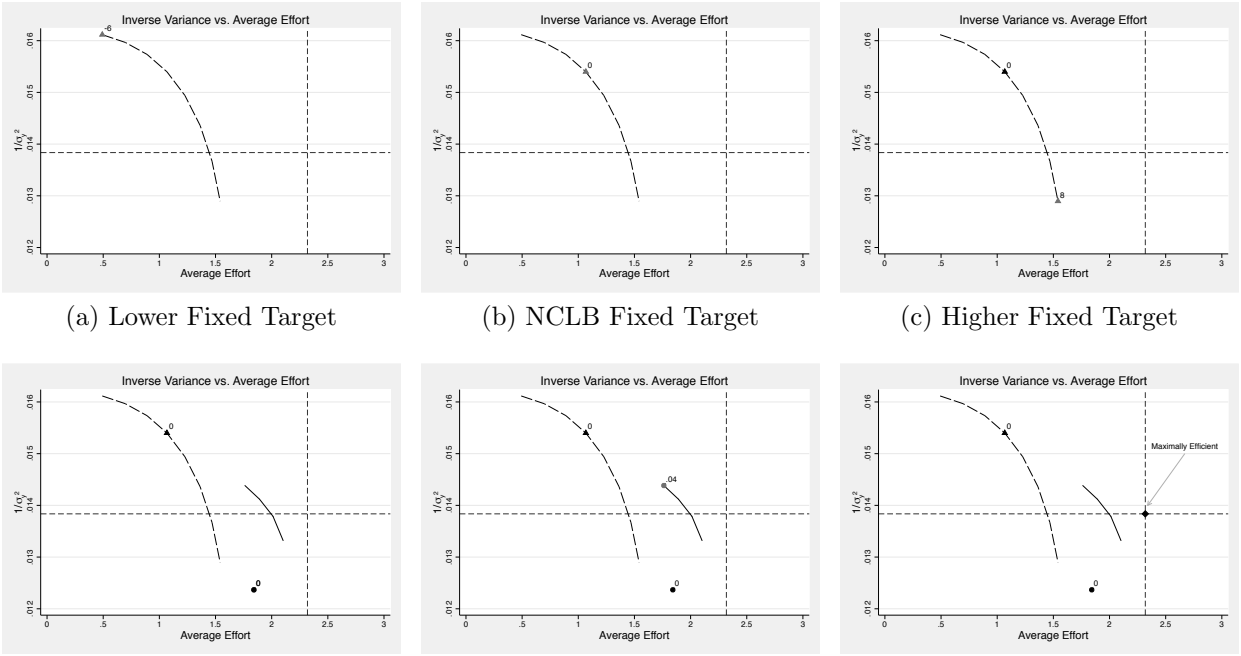
FIGURE 6 – π DENSITIES

We present the results of selected counterfactual simulations in Figure 7. For each simulation, we obtain the two summary measures of interest: average effort across students, and the variance of student test scores. In the analysis below, we plot the *inverse* of the variance against average effort, using the results from all the simulations associated with a given type of scheme to trace out the frontiers for both the set of fixed and VA targets in Average Effort-Inverse Variance space.

The maximally efficient point provides an important benchmark (and is shown in Figure 7f). For each counterfactual regime, we equate the cost to the cost prevailing under the

substantially attenuated in comparison to simpler schemes. Yet it still serves as a useful yardstick in the comparisons below.

maximally efficient target and then recalculate the normalized effort responses, using them to determine final test score outcomes. The normalization allows us to make efficiency comparisons across the regimes by exploring the levels of average effort that prevail for a *given* cost. All points in Figure 7 represent average effort and inverse variance pairs obtained after equating costs across all regimes.



(a) Lower Fixed Target (b) NCLB Fixed Target (c) Higher Fixed Target

(d) ABCs VA Target (e) Higher VA Target (f) Maximally Efficient Point

FIGURE 7 – COUNTERFACTUAL SIMULATIONS

The figure shows there is a clear tradeoff between average effort and inverse variance. To explain the tradeoff for the class of fixed targets, consider the target that is 6 points below the NCLB target in Figure 7a). In this case, virtually the entire π distribution is to the right of the new proficiency threshold, implying that a large fraction of students with relatively high predicted scores fall into the region where the optimal effort function is negative, which works to lower average effort substantially. At the same time, however, the variance is lowest under this regime because only the relatively high-achieving students receive negative effort, while the relatively low-achieving students receive high values of positive effort. The combined effects work to decrease the variance by raising the performance of low-performing students at the expense of high-performing students.

As one shifts the target up, the π distribution begins to shift back to the left, raising

average effort and variance. When the target is 8 points above the NCLB target, the π distribution is virtually centered around zero, implying that there is a fraction of students with large negative values of π and correspondingly low values of effort. The variance under this target is high because students with low predicted scores receive small positive or even negative values of effort and students with relatively high predicted scores receive large positive values of effort. Clearly, the choice of target is critical for determining the amount of effort teachers exert toward each type of student and the variance in scores that results from these effort choices.

Figures 7d), 7e), and 7f) show the frontier for the class of VA proficiency targets. When the multiplicative coefficient is equal to the ABCs coefficient, the resulting point is interior to the VA frontier. This result can be explained as follows: as shown in Figure 6b), the π distribution under the ABCs VA target is quite tight and has most of its mass to the right of zero. Since the peak of the optimal effort function occurs when π is close to zero, the low dispersion of the π distribution makes it possible to substantially raise average effort by only slightly increasing the target and shifting the distribution to the left. When the VA coefficient increases above the ABCs level and the π distribution shifts left, some students fall into the negative effort region of the optimal effort function. In contrast to the fixed scheme, however, these are the students with *high* predicted scores.[45]

Increasing the VA coefficient thus increases the effort most students receive while resulting in negative effort being exerted only toward relatively high-achieving students. The combined effects work to substantially raise average effort while also reducing the variance in student outcomes. As the frontier shows, continuing to increase the VA coefficient eventually causes average effort to fall, as the coefficient that is 0.04 points above the ABCs coefficient results in less effort than that under the ABCs coefficient. It also results in lower variance in tests scores, however, as relatively high-achieving students receive the lowest values of effort. Thus, regular VA schemes also present policymakers with a tradeoff between the variance in outcomes and the average effort exerted.

In terms of comparing the prospective targets, we first note that one can achieve a far

[45]This follows because the coefficient multiplies a student's prior score, implying that large coefficients impose targets above predicted scores for high-achieving students (who also have high prior scores, on average).

lower variance in test scores by using fixed targets rather than VA targets. This follows directly from the relative dispersions in π between the two types of target. Since VA targets are student-specific, the π distribution is much tighter under VA schemes and the corresponding effort responses are more uniform. In contrast, fixed targets generate more dispersion in effort responses and, as a result, can substantially improve outcomes for low-performing students at the expense of high-performing students. Which scheme should be adopted in practice ultimately depends on societal preferences and the tradeoff one is willing to make between variance in outcomes and average effort. The frontiers in Figure 7 imply that only a social planner with a strong aversion to inequality (or a low marginal rate of substitution) would choose a fixed target over a VA target.

Turning to a comparison of specific targets, the NCLB fixed target results in 117 percent more effort and 4 percent higher variance than the lowest fixed target, and it produces 31 percent less effort and 16 percent less variance than the highest target on the fixed frontier. The ABCs target is interior to the VA frontier and results in 72 percent more effort and 25 percent more variance than the NCLB target. Relative to maximally efficient uniform target, which maintains the pre-existing inequality of scores (since all students receive the *same* effort), the ABCs target results in 79 percent of the maximally efficient effort and the NCLB target results in 46 percent of the effort.

Extensions

We further explore our counterfactual analysis along two dimensions, considering school heterogeneity first, followed by the scope for using heterogenous payments (or penalties) in combination with more efficient targets.

School-level analysis

Building on the prior analysis, schools can be viewed as a simple aggregation across students within the school, consistent with a technology that allows teachers to tailor effort to individual students. At the other extreme, each teacher in a grade could be constrained to choose a single level of classroom effort. We conduct the school-level analysis on these two separate bases, providing informative bounds as to the likely effects on the counterfac-

tual performance distribution. The details of this analysis are given in Appendix C: we summarize the main findings here.

Assuming effort is student-specific, 85 percent of variation occurs within schools (and 90 percent of that within classrooms). This accords with the literature. In Kane and Staiger (2002), for example, 87 percent of variance in math scores is within-school; and in Chetty, Friedman and Rockoff (2014), 85 percent of the variance in teacher quality is within-school.

At the other extreme, assuming effort is classroom-specific, the total variance falls by 50-75 percent, and 50 percent of the remainder occurs between schools. (Intuitively, most heterogeneity occurs within-classroom, but that channel is now shut down.)

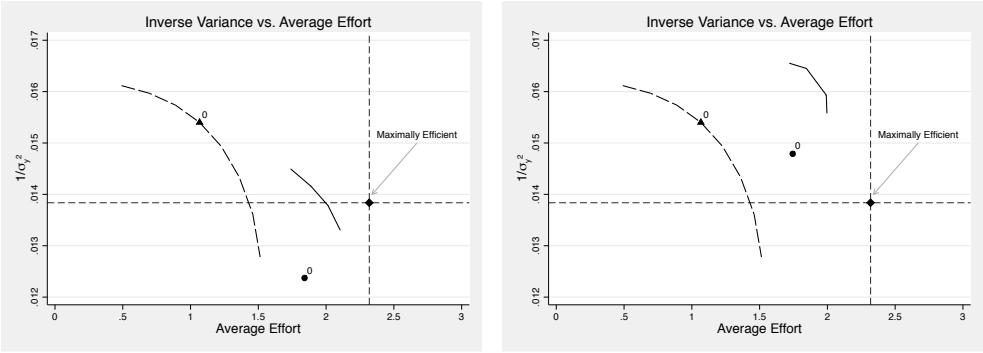
Heterogenous rewards

Thus far, the focus has been on the choice of *target* to determine average effort and the dispersion of outcomes. We have shown that while value-added schemes result in greater effort on average (compared to fixed targets), the scope for altering dispersion is limited. Combining value-added targets with *heterogeneous* bonus payments (or penalties) allows for arbitrary levels of dispersion while preserving average effort. This is highly appealing from a policy perspective, as it opens the possibility of increasing average effort while lowering the variance in outcomes.

We demonstrate the potential for heterogeneous payments as follows: First we divide students into quartiles of the predicted test score distribution, serving as an approximation to varying the reward parameter b continuously. Next we assign a different value of b to each quartile, making three of the quartiles proportional to a reference quartile (e.g. the first quartile). Based on these proportions, we then adjust b for the reference quartile until the total cost across all students is equivalent to the cost under the homogenous b scheme.

For reference, recall that compared to the NCLB target with homogenous rewards/penalties, an ABCs target with a homogeneous reward parameter b results in 20 percent greater variance but also 75 percent greater average effort. We show that, compared to the same NCLB target, an ABCs target with *heterogeneous* rewards achieves nearly identical variance (4 percent more) while attaining 62 percent greater average effort, as Figure 8 illustrates.[46]

[46]In the figure, we set $b_1 = 1.5b_2 = 2.5b_3 = 5b_4$. This arrangement provides higher rewards at the bottom



(a) Homogeneous b (b) Heterogeneous b

FIGURE 8 – FRONTIERS UNDER HOMOGENEOUS AND HETEROGENEOUS b

VI. CONCLUSION

This paper has made two main contributions. First, it offers a transparent semi-parametric approach to identifying the impact of incentives on effort. We used exogenous incentive variation associated with a prominent accountability reform to identify the effort response of North Carolina teachers and schools, showing that the response is consistent with a simple model of effort setting: despite its simplicity, the model fits the data surprisingly well. Our approach is based on minimal assumptions, is easy to implement, and can be applied in other contexts where detailed administrative data are available.

Second, it provides a new framework for measuring the performance of different incentive schemes in education on a comparable basis. The framework incorporates the estimated effort function and allows us to compute the full distribution of scores under counterfactual incentive provisions. To illustrate the power of the framework, we calculated the average effort and dispersion of scores associated with rival incentive schemes, tracing out corresponding performance frontiers on an intuitive, comparable basis.

Among the main findings, our estimates make clear the tradeoff between average effort and the inverse variance of scores for both fixed and value-added targets – the most widespread in education. We show that school heterogeneity is important, the evidence (in Appendix C) suggesting that fixed schemes perform better in schools with a greater pro-

end of the ex ante test score distribution, which decreases its variance ex post and shifts the value-added points in Figure 8a) up to what is seen in Figure 8b). Under the approach, we are able to consider arbitrary relative weights as well as closer and closer approximations to student-specific rewards.

portion of low-performing students. Further, value-added targets with heterogeneous bonus payments across students dominate most fixed targets.

The framework has more general relevance to incentive design issues in organizations. Where sufficiently rich data are available, it provides a tool for measuring the effects on firm output and worker productivity associated with different worker incentive schemes, including schemes that have not yet been implemented. The counterfactual output allows incentive designers to trace the effects of reforms on the dispersion of productivity (with its equity implications) in addition to aggregate productivity effects.

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

References

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

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

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

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

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

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

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

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

Figlio, David N. and Joshua Winicki. 2005. "Food for thought? The effects of school accountability plans on school nutrition." *Journal of Public Economics*, 89(2-3): 381-94.

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

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

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

Lavy, Victor 2009. "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics." *American Economic Review*, 99(5): 1979-2011.

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

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

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

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

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

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

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

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

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

Appendices

A. Contribution of Research Design to Prior Literature

Given the incentive strength measure is very much related to, and builds upon, measures appearing in related prior work, it is worth drawing attention to seemingly subtle differences that turn out to be important in the development of our approach. First, our predicted student scores are based on *pre-reform* data – important in that we use these data to control for baseline effort (described shortly). Similar to the prediction algorithm in Deming, Cohodes, Jennings, Jencks and Lopuch (2013) and Reback (2008), we employ a flexible specification involving lagged test scores and several other student characteristics to calculate expected outcomes, though neither prior study has a pre-reform period.[47] Second, ours is a *continuous* measure, which we can compute for each student. In contrast, Deming *et al.* (2013) aggregate incentive strength to the school-level, and Neal and Schanzenbach (2010) group students into deciles of the ability distribution. The continuous measure is important when conducting counterfactuals, allowing us to evaluate how various targets change incentives throughout the entire student distribution.

B. Ruling Out Rival Effort Story

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

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

inferences about competing accountability schemes would seem unwarranted.

To assess this possibility, we determine the effort responses and corresponding π densities separately for eight types of school. Specifically, we divide schools according to whether they are above or below median variance of π, and further, on the basis of which quartile (in terms of the mean value of π) they are in. If schools responded to NCLB by tailoring effort toward a particular part of the ability distribution, we should observe the peak of the effort response shifting to the right as that point in the ability distribution shifts right across the types of school. This is not the case: Figure A.4 plots the effort responses and π densities separately for schools in each of the quartiles of the mean, focusing on below-median variance.[48] As one moves up the quartiles, the π distribution shifts rightward, implying that a student with a value of π near zero in quartile-one schools will have a different relative position in the $\hat{y}$ distribution than a student with a value of π near zero in the quartile two, three or four schools.

The figure shows that the peak effort response occurs close to $\pi = 0$ and the effort function maintains a similar shape across each of the quartiles. This supports the view that schools respond to a student's proximity to the proficiency threshold and not his or her relative position in the predicted score distribution.

C. Exploring Heterogeneity Across Schools

C.1. Aggregate Differences by School Type

In this subsection we investigate how manipulating the proficiency targets counterfactually affects the distributions of incentive strength (π), effort, and realized scores, both across and within schools. To get a sense of how these quantities evolve as targets change, Figure A.6 groups schools by quartiles of the school-specific mean value of π that prevails under NCLB and then plots changes in the mean (on the top row) and variance (on the bottom) of π, effort, and realized scores for each type of school. In each panel, the horizontal axis measure the distance of the counterfactual fixed target from the NCLB target.

We discuss the average and variance panels in pairs, starting with panels (a) and (d), which plot the evolution of the average and variance of π, respectively. The patterns are straightforward: since the target is constant for all students, the average value of π (predicted score minus fixed target) is linearly decreasing with the target, but changing the (constant) target does not affect

[48] Analogous results hold for schools with above-median variance.

the variance of π. In panel (b), average effort is monotonically increasing at the quartile-four schools. These schools have the best prepared student body, and so higher targets provide sharper incentives for teachers, who continue to exert more effort. At the quartile-one and quartile-two schools, however, average effort eventually begins to decline. Students are poorly prepared at these schools, and increasing the target reduces the likelihood of many students reaching proficiency status, resulting in weaker incentives for teachers to exert effort.

Panel (e) shows each school type has a target level for which the variance is minimized before starting to increase rapidly at higher values of the target. The target corresponding to the minimum variance is increasing in the quartile to which a school belongs, with the minimum achieved at 0, 2, 4, and 6 points above the NCLB target across quartile-one, two, three, and four schools, respectively. This result is explained by the shape of the effort function and the distribution of π across each type of school. For values of π between 0 and 10, the effort function is relatively flat, assigning fairly uniform levels of effort to each value of incentive strength. Intuitively, the fixed target that results in the largest mass of the π distribution being contained in the uniform-effort range also results in the minimum effort variance. Since the π distribution mechanically shifts right as school quartiles increase, it takes a progressively higher value of the fixed target to pull the distribution back into the range where effort is uniform and the variance is minimized.

As one would expect from the profiles of average effort in panel (b), panel (c) shows that average test scores are increasing with the target at quartile-four schools, begin to level off at quartile-three schools, and begin to slightly decline at high target values at quartile-two and quartile-one schools. The variance of test scores in panel (f) is increasing with the target at all types of school. At higher target values, the low-performing students at each school receive little (or negative) effort and the high-performing students receive high effort, which works to exacerbate inequality, driving up the variance in realized scores.[49]

Figure A.7 shows the relevant patterns for the class of VA targets. The horizontal axes measure the distance of the counterfactual VA multiplicative coefficient from the ABCs coefficient (0.68).[50] A key insight from Figure A.7 is that there is much less (in fact, almost zero) heterogeneity in the

[49]Note that the variance of test scores does not follow the same profile as the variance of effort. While the counterfactual test score is the sum of predicted scores, noise, and effort, the variance of test scores is not the sum of the variances of these three components. Effort is a function of the predicted score, implying a non-zero covariance component between the two.

[50] Note that, because North Carolina's ABCs program uses coefficients and targets exclusively with test scores measured on the first-edition scale in grade four, all of the results for VA targets are measured on the first-edition scale. This is in contrast to the results above for the fixed targets, which are measured on the second-edition scale.

average value of π across school types. VA schemes effectively set student-specific targets, implying that incentive strength is similar across all students. Panel (d) shows the variance of π is increasing in the VA coefficient, implying that higher values of the coefficient shift the distribution of π to the left while increasing its spread. Despite the variances increasing, they are still much lower than the variances that prevail under fixed targets.

In panel (b), average effort is steadily increasing in the VA target until beginning to fall at a coefficient 0.02 points higher than the ABCs coefficient. At this point, the π distribution shifts far to the left and a large mass of students at each school begin receiving negative effort, which lowers the average. Test scores follow a similar profile in panel (c), slightly increasing and then slightly decreasing for high values of the coefficient.

As mentioned, the effort function is relatively flat for most values of π between 0 and 10, and then exhibits an abrupt decline once π is lower than -2. For values of the multiplicative coefficient ranging between -0.04 and 0.01, the π distribution has a relatively small variance and falls in the region where the effort function is quite flat. Panel (e) thus shows that there is virtually no variance in effort for most values of the coefficient. At high values of the coefficient, the π distribution shifts far to the left and its variance increases, resulting in many students falling into the domain where effort abruptly declines and becomes negative for students with the lowest values of π. This works to increase the dispersion in effort across students at each type of school sharply. In panel (f), one can see these relations between π and effort cause a sudden reduction in the variance of tests scores. Under the VA scheme, students who are pushed into the negative effort region by high multiplicative coefficients are the high-performing students, implying that the reduction in variance is achieved by lowering the achievement of the historically high-performing students.

C.2. Within- and Between-School Variances

In this subsection, we document the implications of the dynamics discussed above for the inequality of student outcomes within and across the full set of schools, no longer grouping schools into four types. Under each counterfactual target, Table A.2 decomposes the variance of π, effort, and realized scores into the variance within schools, the variance between schools, and the amount of the within-schools variance that occurs within classrooms.

Most of the variation in each variable occurs within schools, which is line with several studies that find much of the variance in education variables occurs within, not across, schools (see, for example, Kane and Staiger, 2002). While there is some change in the ratio of the within-school

variance to total variance as one changes the targets, these changes are very small. For example, a fixed target six points below the NCLB target results in 88 percent of the variance in total test scores occurring within schools, while a target ten points above NCLB results in 86 percent of the variance occurring within schools. In absolute terms, however, higher fixed targets result in much higher inequality in final outcomes, both within and across schools: the within-school variance in test scores is 33 percent higher under the fixed target ten points above NCLB than the target six points below NCLB; the between-school variance is 53 percent higher.

While higher fixed targets increase the variance of tests scores, higher VA coefficients decrease it. Panel B of Table A.2 shows that opposite patterns prevail when one moves from a VA coefficient 0.04 points below the ABCs coefficient to one 0.04 points above. The total variance in test scores declines by 14 percent, and the within- and between-schools variances fall by 15 and 14 percent, respectively.

Under both classes of scheme, about 90 percent of the within-schools variation in each variable occurs within classrooms. To some extent, one might be concerned that this is an artefact of the assumption that teachers can choose to adjust their effort across students on a student-specific basis in a flexible way. Yet the evidence actually supports this assumption.[51] This lends confidence to our counterfactual results.

While we do not know the exact combination between student- and classroom-specific effort teachers may choose, we can create bounds for our results by assuming that each student in a classroom receives the *same* level of effort, and then conducting the counterfactual analyses while maintaining that assumption. We can thus consider the extreme cases of student- and classroom-specific effort separately, while knowing that the true data-generating process likely lies somewhere in between the two. The following subsection reports results under the assumptions that each student in a classroom receives the same amount of effort and that teachers reach decisions about classroom effort levels according to the model in Section I.A.

[51] This is because we do not use the model to estimate the actual effort response under NCLB. Rather, that is simply the realized minus predicted score for each student, and there is nothing structural about it. Yet in Table A.2, we see that even under the 0 column in panel A (which reflects what actually happened), most of the variance is within schools and within classrooms. This suggests that we are closer to the student-by-student world than the common-effort-within-classrooms world.

C.3. Classroom-Specific Effort Constraint

Table A.3 shows how the variances of each variable evolve when effort is constrained to be equal within classrooms. As in Section I.A, teachers choose the level of effort corresponding to the average values of π within their classrooms. Since we do not change anything with respect to the student-specific incentive strength, π, the variances of π are the same as those in Table A.2. The within-classroom variance of effort is zero by construction under the assumption of common classroom effort, and the within-classroom variance in test scores is constant across the targets within the fixed and VA schemes.[52]

Under the common effort assumption, the between-school variance becomes much more important in explaining the total variation in effort, comprising 40 to 60 percent of the total effort variance within the class of fixed targets and about 50 percent of the variance within the class of VA targets (the last two columns of panel B). Much of the dispersion in test scores still occurs within schools, however, the within-classroom variance explaining less of the within-school score variance for higher fixed targets and more for higher VA coefficients.

In the prior subsection and this one, we have considered the distributional implications of two extreme cases: one where all effort is student-specific and the other where effort is constrained to be uniform within classrooms. In practice, the effort exerted by teachers is likely to fall somewhere in between. In the absence of weights to determine the importance of each case, the comparison is still useful, as it helps us bound the actual effect of varying the incentive target counterfactually.

C.4. Differential Effort Functions Across School Types

While teachers are likely constrained in the extent to which they can differentiate effort across students within a classroom, this may not be the only source of heterogeneity in optimal effort responses. In particular, two students who have the same level of incentive strength, π, but attend different schools may receive different levels of teacher effort, depending on the likelihood that their schools satisfy the standard of the accountability program in question.[53] It is plausible to think that the amount of effort each student receives depends both on his or her individual likelihood of passing and the probability that his or her school passes: if the school is very likely (or unlikely)

[52]They are not constant across the schemes because the VA results use the first-edition math scale and the Fixed results use the second-edition. See footnote 50.

[53]For example, a student on the margin of proficiency in a school with a reasonable likelihood of passing the standard may receive a large amount of additional effort while a similar student in a school with a very small (or high) chance of passing may receive no additional effort.

to pass no matter the actions it takes, it may not respond to the accountability provisions at all.

We already explored this type of of heterogeneity to some extent in Figure A.4 in Appendix B, where we divide schools by their distributions of π and look for heterogeneous effort responses across types of school. There, we find little evidence that schools with a lower (first quartile schools) or higher probability (fourth quartile schools) of doing well under NCLB responded by targeting effort differently toward students. Instead, all types of school seem to be responding to the proximity of each student to the proficiency target rather than a school-specific probability of passing. Nevertheless, we allow for differential effort responses by school type in this subsection by recalculating the optimal effort function within each school type and redoing the counterfactual analyses using these school-type-specific effort functions. To explore the contrast between schools with low- and high-ability students, we consider schools with a variance of π below the median and present separate results for those with mean values of π in the first and fourth quartile.

Low-Mean, Low-Variance Schools

Figure A.8 shows the fixed and VA target frontiers for schools with a mean value of π in the first quartile. Students at these schools are predicted to have relatively low performance in the absence of any additional effort. There still exists a clear tradeoff between the variance in test scores and the average effort exerted within both the set of fixed and VA targets. The NCLB target continues to balance average effort and the dispersion of scores relative to other fixed targets, as it achieves 50 percent more effort and 6 percent higher variance than the lowest fixed target, and it produces 2 percent less effort and 20 percent less variance than the fixed target that is 8 points higher than NCLB.

Unlike the aggregate results, the 8-point-higher fixed target is interior to the fixed frontier among schools serving low-performing students. Since many students at these schools are predicted to have low scores, a target this high makes it very difficult for them to reach proficiency status. Teachers recognize this and opt not to direct resources toward these students, causing their test scores to fall and score inequality to be exacerbated.

The VA targets result in more effort and inequality than the fixed targets. For example, the ABCs VA target results in 28 percent more effort and 30 percent more inequality than the NCLB fixed target. For schools serving low-performing students, the highest variance in test scores generated by the fixed targets is lower than the variance in test scores produced by all-but-one VA target (the VA target when the coefficient is 0.04). Since inequality is so much lower under the

set of fixed schemes, even a planner with a moderate aversion to variance in scores might choose a fixed target for these schools.

These reductions in inequality come at the expense of the high-performing students in these schools, however. As there are relatively few of those students, when faced with a fixed target, teachers in such schools really focus on redirecting their attention toward students near the proficiency threshold, choosing large negative values of optimal effort for the highest-performing students. When one lowers the fixed target and pushes the π distribution to the right, an increasing fraction of relatively high-performing students are shifted into the area where effort is negative, leading to a substantial decline in inequality.

High-Mean, Low-Variance Schools

Figure A.9 shows the fixed and VA target frontiers for schools with a mean value of π in the fourth quartile. Students at these schools are predicted to perform highly in the absence of any additional effort. In such schools, the VA frontier clearly dominates the fixed frontier, as fixed targets cannot produce the same reduction in variance as they can in the schools that serve low-performing students.

Since high performers represent a relatively high fraction of students in these schools, teachers do not redirect resources away from them to the same degree that they do in low-performing schools, which results in an effort function with small negative values of effort for large values of π and a relatively flat profile for progressively larger values of π. When one lowers the fixed target and pushes the π distribution to the right, the high-performing students pushed into the area where effort is negative only experience small declines in performance, resulting in relatively small changes in inequality. In high-performing schools, VA schemes clearly dominate and would be chosen by a planner, regardless of preference over the variance of scores and average effort.

Table A.1 – Descriptive Statistics

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

School-Level			
	Mean	*Std. Dev.*	*N*
Failed ABCs	0.27	0.45	14,052
Failed NCLB	0.37	0.48	5,014
Proficiency Rate	0.79	0.11	14,042

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

Table A.2 – Variance Decomposition Across Counterfactual Regimes with Student-Specific Effort

	A: Fixed Target Relative to NCLB								
	−6	−4	−2	0	2	4	6	8	10
Total Variance π	49.09	49.09	49.09	49.09	49.09	49.09	49.09	49.09	49.09
Between School	9.16	9.16	9.16	9.16	9.16	9.16	9.16	9.16	9.16
Within School	39.93	39.93	39.93	39.93	39.93	39.93	39.93	39.93	39.93
Within Class	35.6	35.6	35.6	35.6	35.6	35.6	35.6	35.6	35.6
Total Variance Effort	0.63	0.57	0.49	0.4	0.34	0.32	0.42	0.7	1.28
Between School	0.11	0.1	0.09	0.06	0.04	0.02	0.02	0.05	0.14
Within School	0.51	0.46	0.4	0.34	0.29	0.3	0.4	0.65	1.14
Within Class	0.46	0.42	0.36	0.3	0.27	0.27	0.37	0.6	1.04
Total Variance Score	62.05	62.63	63.55	64.93	66.95	69.74	73.47	78.25	84.36
Between School	7.55	7.65	7.81	8.06	8.41	8.91	9.59	10.46	11.56
Within School	54.5	54.98	55.73	56.87	58.53	60.83	63.88	67.79	72.8
Within Class	49.99	50.41	51.08	52.09	53.56	55.61	58.33	61.83	66.32

	B: VA Coefficient Relative to ABCs								
	−0.04	−0.03	−0.02	−0.01	0	0.01	0.02	0.03	0.04
Total Variance π	2.01	2.19	2.39	2.61	2.85	3.12	3.4	3.71	4.04
Between School	0.12	0.14	0.15	0.18	0.2	0.23	0.26	0.3	.33
Within School	1.89	2.05	2.23	2.44	2.65	2.89	3.14	3.41	3.7
Within Class	1.75	1.9	2.07	2.25	2.44	2.66	2.88	3.13	3.39
Total Variance Effort	0.01	0	0	0	0	0.01	0.15	0.71	1.44
Between School	0	0	0	0	0	0	0.01	0.05	0.1
Within School	0.01	0	0	0	0	0.01	0.14	0.66	1.33
Within Class	0.01	0	0	0	0	0.01	0.13	0.61	1.23
Total Variance Score	80.97	80.75	80.63	80.72	80.83	80.23	77.59	72.55	69
Between School	10.78	10.77	10.75	10.76	10.78	10.71	10.39	9.71	9.23
Within School	70.19	69.98	69.88	69.95	70.05	69.53	67.2	62.84	59.77
Within Class	63.98	63.8	63.7	63.77	63.86	63.4	61.33	57.46	54.72

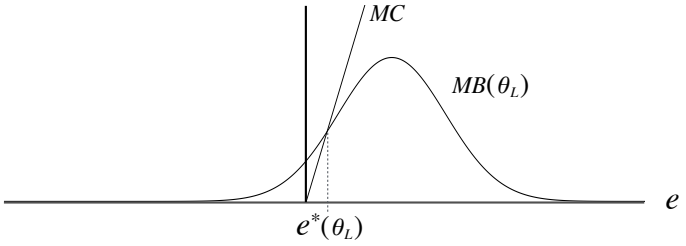
Notes: The total variance of a given variable is calculated across all students. For each variable, the within- and between-school variances decompose the total variance into the variance that occurs within schools and the variance that occurs across schools, respectively. The within-class variance represents the amount of the within-school variance that occurs within classrooms. All available fourth grade students, schools, and classrooms in 2003 are used in the calculations.

Table A.3 – Variance Decomposition Across Counterfactual Regimes with Classroom-Specific Effort

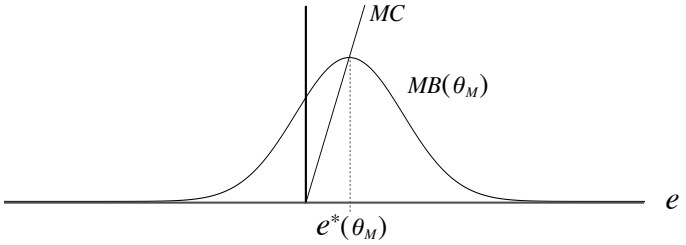
	A: Fixed Target Relative to NCLB								
	−6	−4	−2	0	2	4	6	8	10
Total Variance π	49.09	49.09	49.09	49.09	49.09	49.09	49.09	49.09	49.09
Between School	9.16	9.16	9.16	9.16	9.16	9.16	9.16	9.16	9.16
Within School	39.93	39.93	39.93	39.93	39.93	39.93	39.93	39.93	39.93
Within Class	35.6	35.6	35.6	35.6	35.6	35.6	35.6	35.6	35.6
Total Variance Effort	0.22	0.18	0.14	0.1	0.07	0.05	0.05	0.15	0.49
Between School	0.15	0.12	0.09	0.07	0.04	0.03	0.02	0.07	0.3
Within School	0.07	0.06	0.05	0.03	0.03	0.02	0.03	0.08	0.18
Within Class	0	0	0	0	0	0	0	0	0
Total Variance Score	69.8	70.08	70.43	70.86	71.36	71.88	72.57	74.06	76.77
Between School	7.36	7.56	7.79	8.08	8.39	8.71	9.12	10.11	12.04
Within School	62.44	62.52	62.63	62.78	62.97	63.18	63.45	63.95	64.73
Within Class	58.02	58.02	58.02	58.02	58.02	58.02	58.02	58.02	58.02

	B: VA Coefficient Relative to ABCs								
	−0.04	−0.03	−0.02	−0.01	0	0.01	0.02	0.03	0.04
Total Variance π	2.01	2.19	2.39	2.61	2.85	3.12	3.4	3.71	4.04
Between School	0.12	0.14	0.15	0.18	0.2	0.23	0.26	0.3	0.33
Within School	1.89	2.05	2.23	2.44	2.65	2.89	3.14	3.41	3.7
Within Class	1.75	1.9	2.07	2.25	2.44	2.66	2.88	3.13	3.39
Total Variance Effort	0	0	0	0	0	0	0.01	0.18	0.61
Between School	0	0	0	0	0	0	0	0.09	0.33
Within School	0	0	0	0	0	0	0.01	0.09	0.28
Within Class	0	0	0	0	0	0	0	0	0
Total Variance Score	81.06	81.01	80.98	81.02	81.11	81.07	80.54	78.79	77.07
Between School	10.78	10.76	10.75	10.77	10.82	10.81	10.52	9.51	8.32
Within School	70.27	70.24	70.23	70.25	70.29	70.26	70.02	69.29	68.75
Within Class	64.06	64.06	64.06	64.06	64.06	64.06	64.06	64.06	64.06

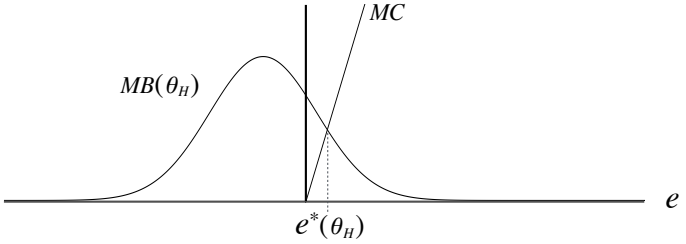
Notes: The total variance of a given variable is calculated across all students. For each variable the within- and between-school variances decompose the total variance into the variance that occurs within schools and the variance that occurs across schools respectively. The within-class variance represents the amount of the within-school variance that occurs within classrooms. All available fourth grade students schools and classrooms in 2003 are used in the calculations.



(a) Low θ relative to y^T



(b) Intermediate θ relative to y^T



(c) High θ relative to y^T

Figure A.1 – Optimal Effort and Varying Exogenous Production Conditions under a Threshold Scheme

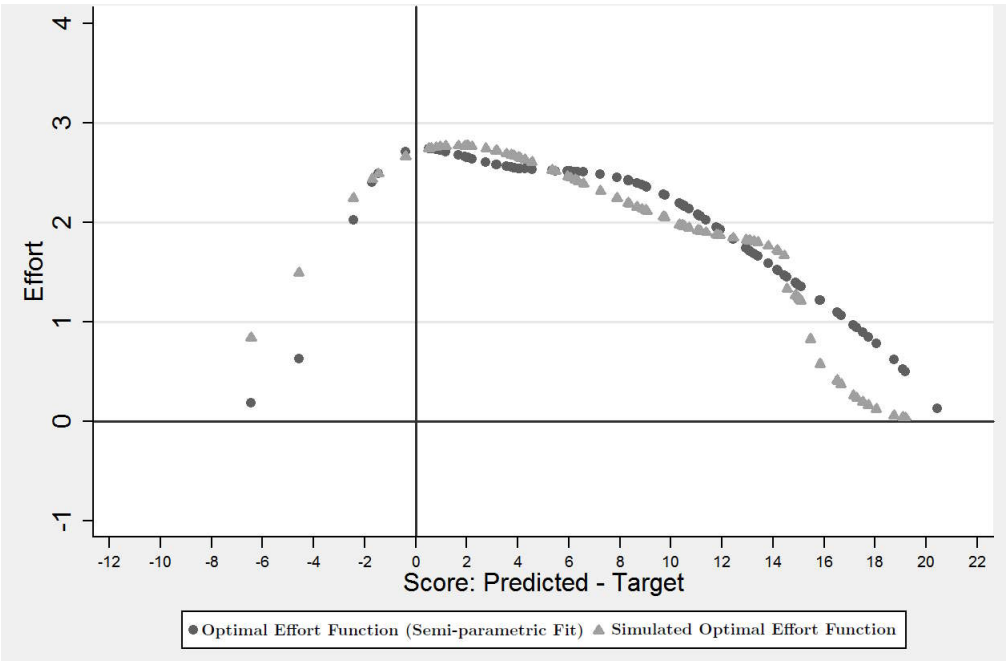
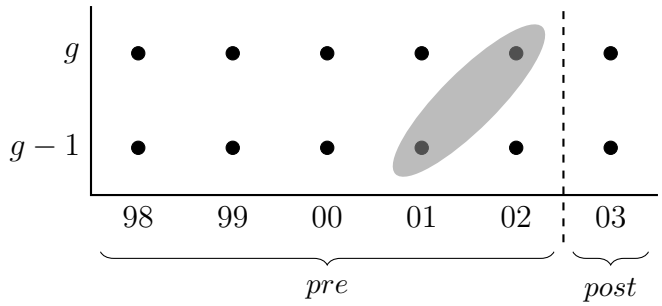
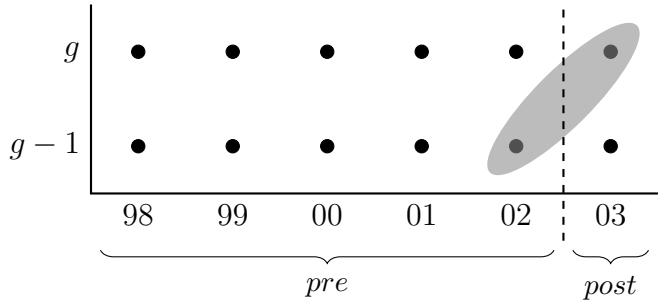


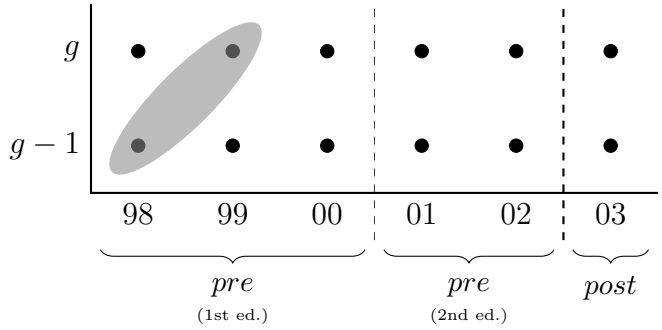
Figure A.2 – Simulated and Semi-Parametric Optimal Effort Functions



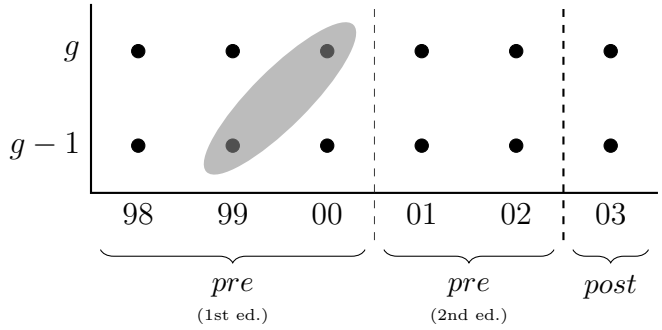
(a) First Step



(b) Second Step

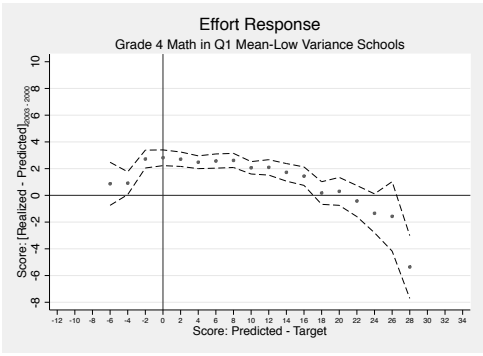


(c) Third Step

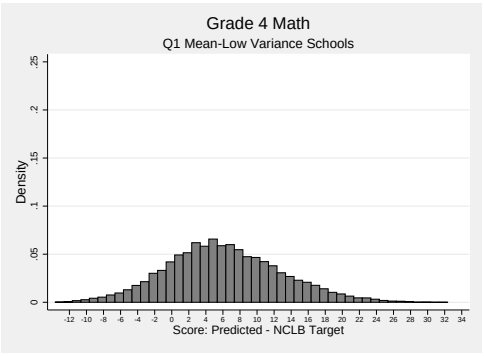


(d) Fourth Step

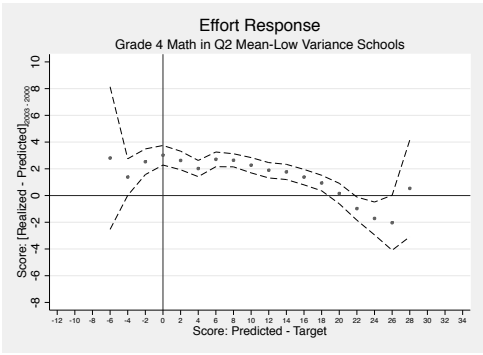
Figure A.3 – Research Design in Pictures



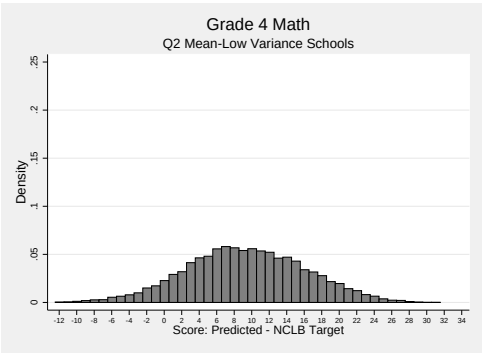
(a) Effort in Q1 Mean Schools



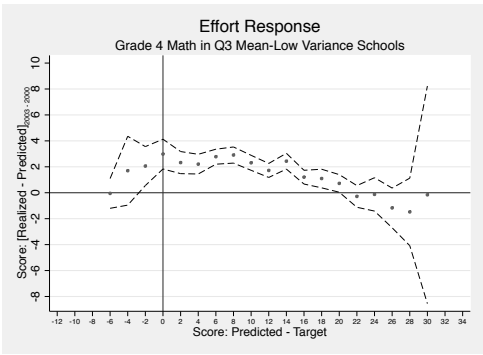
(b) π Density in Q1 Mean Schools



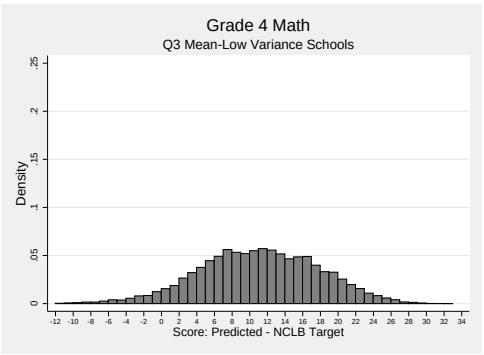
(c) Effort in Q2 Mean Schools



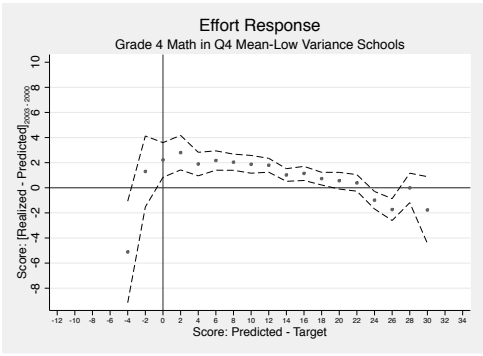
(d) π Density in Q2 Mean Schools



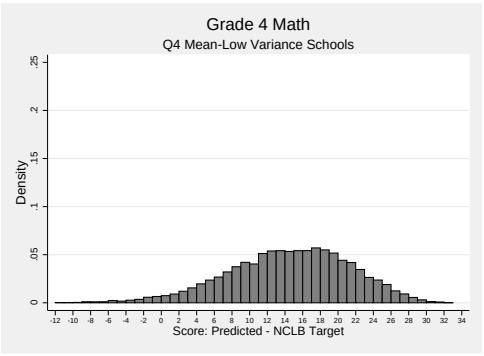
(e) Effort in Q3 Mean Schools



(f) π Density in Q3 Mean Schools

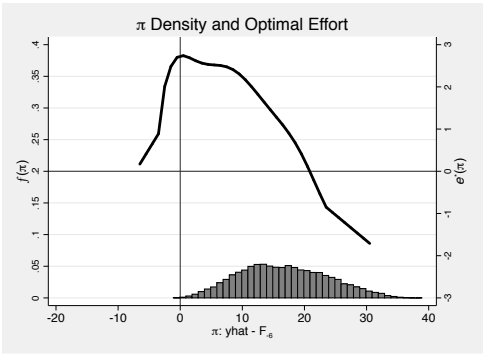


(g) Effort in Q4 Mean Schools

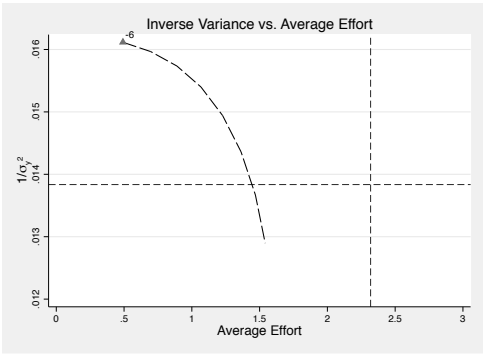


(h) π Density in Q4 Mean Schools

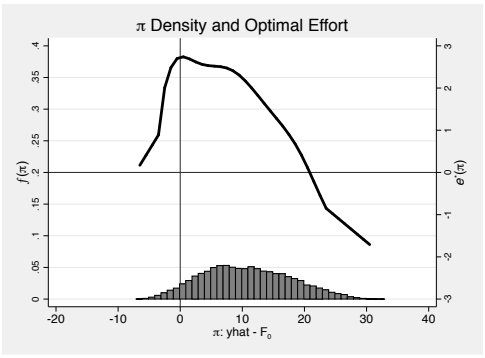
FIGURE A.4 – RESPONDING TO π NOT THE RELATIVE POSITION OF $\hat{y}$?



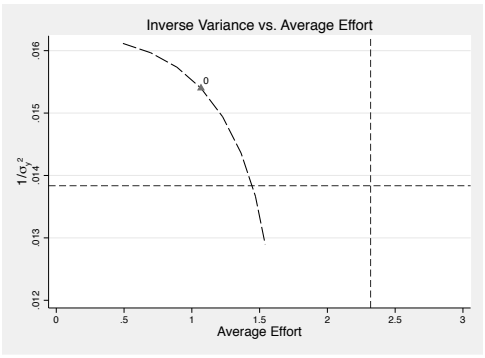
(a) Lower Fixed Target: π and e^*



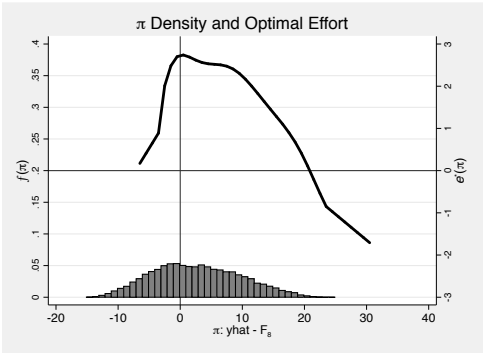
(b) Lower Fixed Target: Implied Point



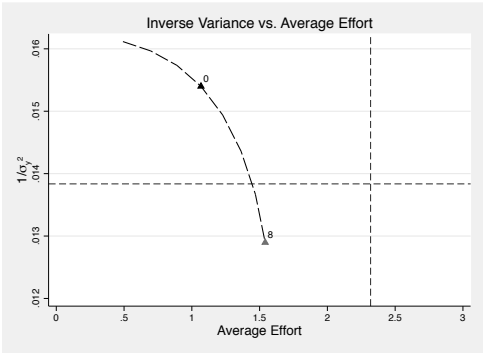
(c) Actual Fixed Target: π and e^*



(d) Actual Fixed Target: Implied Point

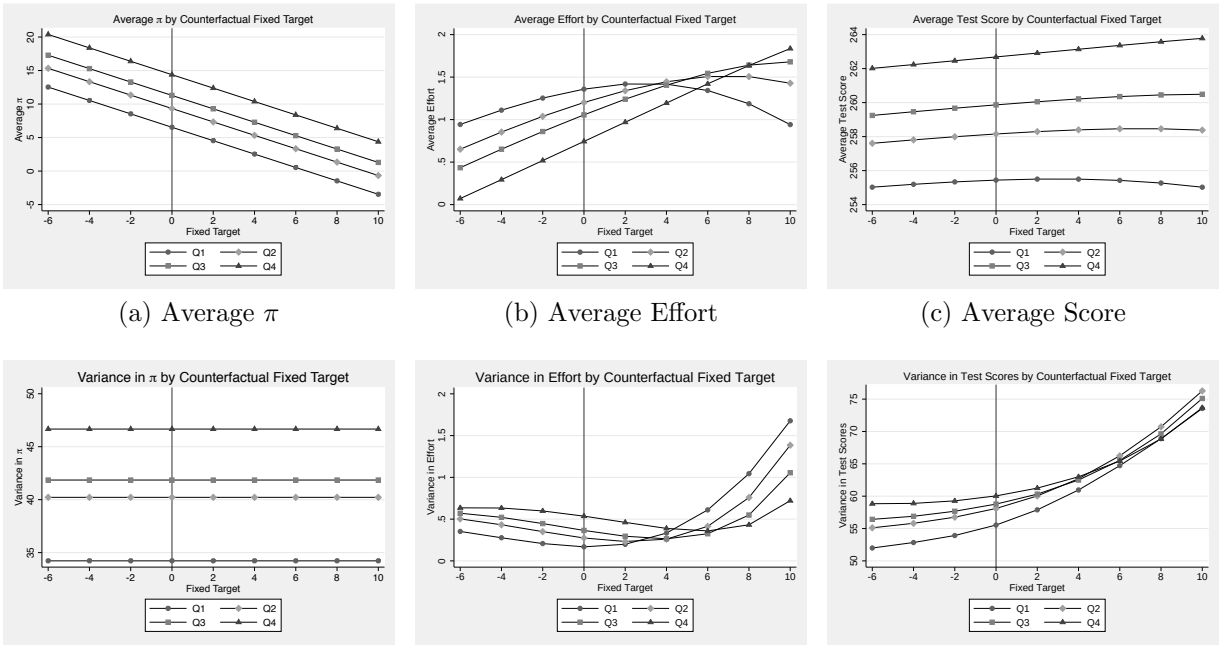


(e) Higher Fixed Target: π and e^*



(f) Higher Fixed Target: Implied Point

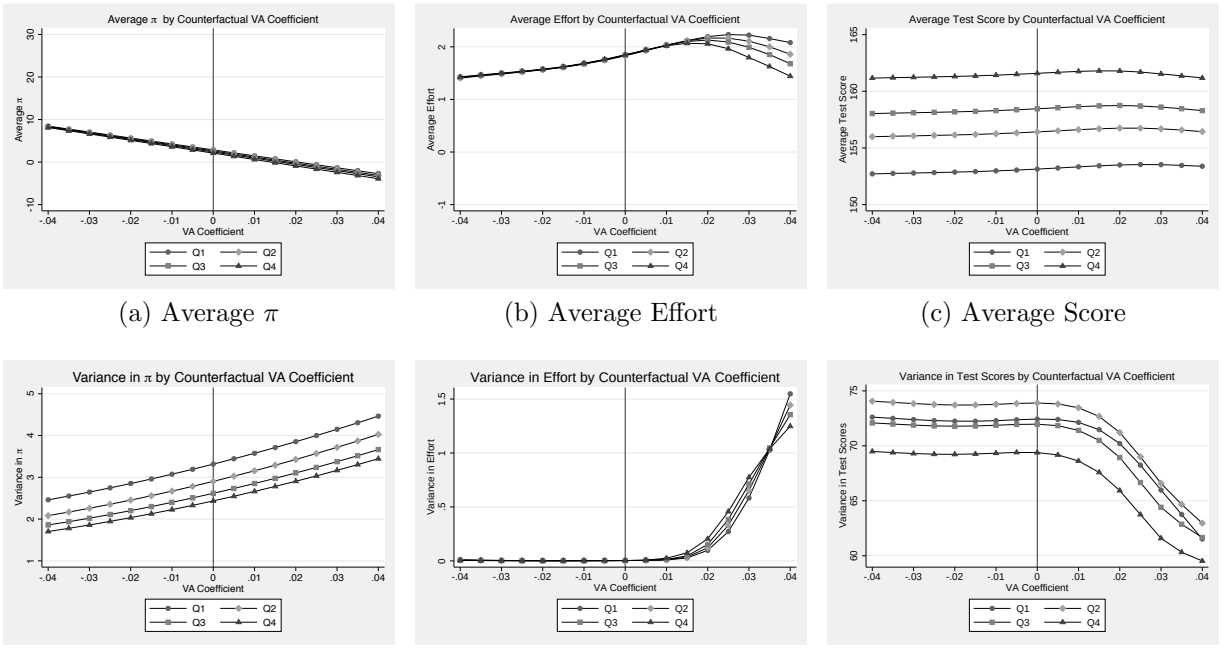
FIGURE A.5 – DERIVING THE FIXED FRONTIER



(a) Average π (b) Average Effort (c) Average Score

(d) Variance of π (e) Variance of Effort (f) Variance of Score

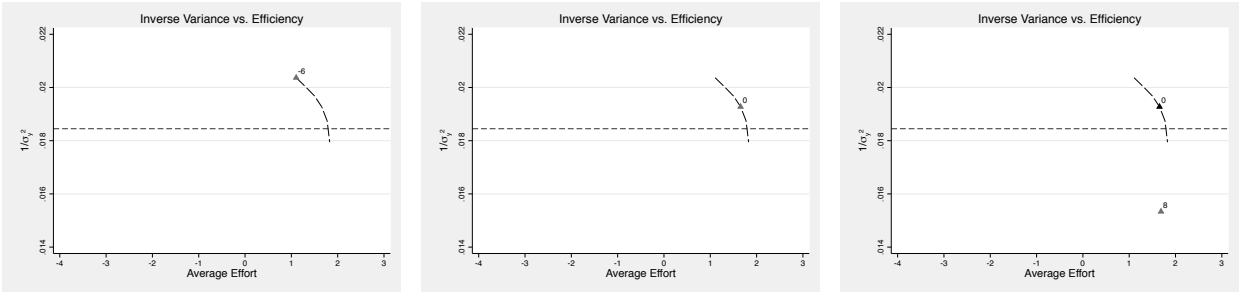
FIGURE A.6 – COUNTERFACTUAL FIXED TARGETS AND SCHOOL-TYPE HETEROGENEITY WITH STUDENT-SPECIFIC EFFORT



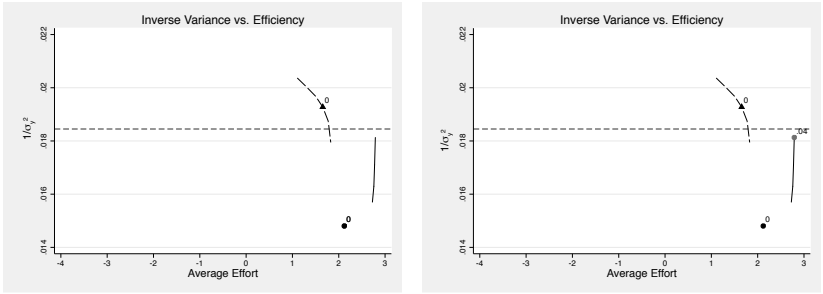
(a) Average π (b) Average Effort (c) Average Score

(d) Variance of π (e) Variance of Effort (f) Variance of Score

FIGURE A.7 – COUNTERFACTUAL VA TARGETS AND SCHOOL-TYPE HETEROGENEITY WITH STUDENT-SPECIFIC EFFORT

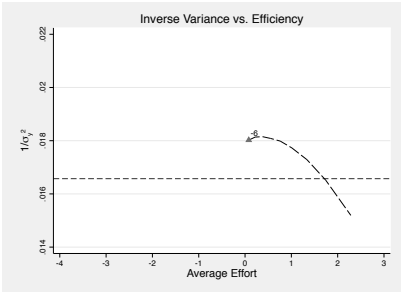


(a) Lower Fixed Target (b) NCLB Fixed Target (c) Higher Fixed Target

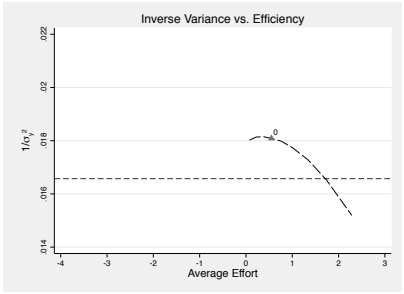


(d) ABCs VA Target (e) Higher VA Target

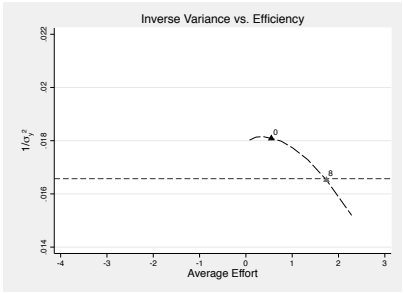
FIGURE A.8 – COUNTERFACTUAL SIMULATIONS FOR LOW MEAN, LOW VARIANCE SCHOOLS



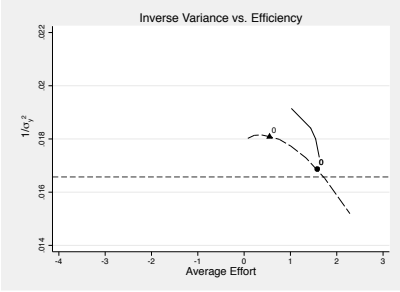
(a) Lower Fixed Target



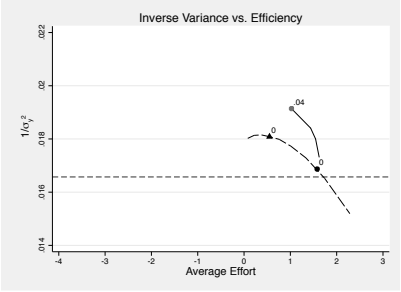
(b) NCLB Fixed Target



(c) Higher Fixed Target



(d) ABCs VA Target



(e) Higher VA Target

FIGURE A.9 – COUNTERFACTUAL SIMULATIONS FOR HIGH MEAN, LOW VARIANCE SCHOOLS