Incentive Design in Education: An Empirical Analysis*

Hugh Macartney† Robert McMillan‡ Uros Petronijevic§

April 13, 2015

*We would like to thank Joseph Altonji, Peter Arcidiacono, Lisa Kahn, Lance Lochner, Aloysius Siow, and seminar participants at Duke University, Western, and Yale University for helpful comments and suggestions. Financial support from the University of Toronto is gratefully acknowledged. All remaining errors are our own.

†Department of Economics, Duke University, 213 Social Sciences Building, Box 90097, Durham, NC 27708, and NBER. Email: hugh.macartney@duke.edu

‡Department of Economics, University of Toronto, 150 St. George Street, Toronto, ON M5S 3G7, Canada, and NBER. Email: mcmillan@chass.utoronto.ca

§Department of Economics, University of Toronto, 150 St. George Street, Toronto, ON M5S 3G7, Canada. Email: uros.petronijevic@utoronto.ca

Abstract

This paper provides the first empirical analysis assessing the cost effectiveness of alternative education accountability systems, including those yet to be implemented. We set out a semi-parametric approach that uses exogenous variation in incentives to isolate the underlying effort response of teachers and schools. This response in hand, we compute the average effort and dispersion of outcomes for a given cost under different accountability systems, including the most widespread – those setting fixed and value-added targets. Varying the parameters of each type of scheme counterfactually allows us, in turn, to trace out associated performance frontiers on a comparable basis. To implement the approach, we use rich administrative data covering all public school students in North Carolina, showing that the introduction of the federal No Child Left Behind Act on top of the state's pre-existing accountability scheme led to marked improvements in scores, with students closer to the margin of passing experiencing larger gains. Differencing the score distributions post- versus pre-reform with respect to a continuous incentive strength measure uncovers the effort function used in our counterfactual approach. The types of estimate to emerge from this are new to the literature: We show that there is a clear tradeoff between average performance and the tightness of the score distribution for both fixed and value-added schemes, and that the value-added frontier lies to the right of its fixed counterpart, associated with higher effort for a given cost. Further, a scheme that makes bonus payments student-specific can, for a given average effort level, reduce the variance of scores below that attainable under a regular value-added scheme. The analysis is relevant to the reform of existing accountability systems, with broader implications for using policy variables to improve education productivity.

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

1 Introduction

As a response to chronic under-performance, accountability schemes have become an integral feature of a host of modern-day public education systems. In their essence, they involve setting performance targets and explicit rewards (or penalties) that depend on target attainment, their goal being to increase the effort of teachers and schools through heightened incentives, thereby raising student test scores.[1] Persuasive evidence that accountability schemes succeed in improving student achievement already exists;[2] further, as noted by Hoxby (2002), they are likely to do so at a cost far lower than reforms – particularly class size reduction – that simply increase inputs.

While sharing a common effort-raising goal, several types of accountability scheme have been implemented to date. These range from proficiency schemes, most notably the federal No Child Left Behind Act of 2001 (henceforth 'NCLB'), setting performance targets based on school sociodemographics, to value-added schemes operating in several states,[3] whose targets condition on prior student scores. This variety brings to mind important incentive design issues, not least the way that changing incentives affects the distribution of student achievement. In particular, do modifications to existing schemes give rise to a tradeoff between average effort and the spread of outcomes, controlling for cost? And could gains be had by moving to other classes of scheme, including those yet to be implemented?

These issues have yet to be addressed empirically. Although several convincing studies consider the design of incentives in education from an empirical standpoint, focusing on particular aspects of schemes already in operation,[4] no prior empirical research assesses the relative merits of rival schemes in a quantifiable way. At minimum, this involves tracing out their effects on the performance distribution on a common footing, based on credibly-identified parameters and a tractable framework for counterfactual analysis.

To help fill this gap, we carry out the first empirical analysis of the impacts of alternative education accountability systems on the full distribution of test scores. We do so by exploiting plausibly exogenous incentive variation arising from the introduction of NCLB. Being a

[1] Measuring teacher effects more generally has been the focus of important recent empirical work, including Rivkin, Hanushek and Kain (2005) and Chetty, Friedman and Rockoff (2014).

[2] See Carnoy and Loeb (2002), Lavy (2002, 2009), Hanushek and Raymond (2005), Figlio and Kenny (2007), Dee and Jacob (2011), and Muralidharan and Sundararaman (2011), among others.

[3] These include Arizona, Colorado, Florida, North Carolina, South Carolina and Texas.

[4] See Kane and Staiger (2002), Cullen and Reback (2006), Neal and Schanzenbach (2010), and Macartney (2014). For example, Cullen and Reback (2006) examine the effect of exempting disadvantaged students from testing; and Macartney (2014) explores dynamic distortions that arise under growth-oriented schemes, focusing on North Carolina.

proficiency scheme, NCLB creates incentives to focus on students at the margin of passing (relative to a fixed target),[5] such non-uniformity providing useful identifying power. We take advantage of this power in North Carolina, a setting for which we have rich administrative data covering all public school students over a number of years, where the NCLB system was implemented on top of the pre-existing statewide 'ABCs,' a relatively uniform growth-oriented accountability scheme.

To guide our empirical analysis, we specify a model of the education process that links incentives to outcomes via discretionary action – 'effort'[6] – on the part of teachers and schools. Student performance is determined by a production technology that depends on such effort, along with exogenous characteristics including student ability, and a noise component. The effort choice of educators is treated as being endogenous to the prevailing incentive scheme, with agents balancing greater rewards against the convex cost of effort. This implies an effort function that depends on the parameters of the incentive scheme and, under threshold targets, a measure of incentive strength describing how marginal the educator is given exogenous circumstances, as is quite standard.

This effort function provides the basis for our empirical analysis. We begin by constructing a new incentive strength measure using the rich North Carolina data. This is a continuous *ex ante* measure that captures how marginal each student is relative to a fixed proficiency target. It is equal to the gap between the target and the student's predicted score, based on data from the pre-reform period.

The measure is very much related to, and builds upon, measures appearing in related prior work, so it is worth drawing attention to seemingly subtle differences that turn out to be important in the development of our approach. First, the predicted student score uses *pre-reform* data, using a flexible specification involving lagged test scores and several other student characteristics to calculate expected outcomes. Our prediction algorithm is similar to that used by Deming, Cohodes, Jennings, Jencks and Lopuch (2013) in their analysis of the Texas accountability program that operated throughout the 1990's, though they do not have a pre-reform period; Reback (2008) calculates a student-level measure of incentive strength – a passing probability rather than a predicted score – using Texas data, though also without a pre-reform period. The pre-reform data are important in that we use them

[5]That NCLB creates such incentives has been well-documented in the literature – see Reback (2008), Ladd and Lauen (2010), and Neal and Schanzenbach (2010), for example.

[6]This labeling is standard in incentive theory – see e.g. Laffont and Tirole (1993). In our setting, 'effort' refers to changes in observable test scores that are attributable to incentive variation.

to control for baseline effort, described shortly. Second, ours is a *continuous* measure, which we can compute for each student, while Neal and Schanzenbach (2010) group students into deciles of the ability distribution and Deming *et al.* (2013) aggregate incentive strength to the school-level. The continuous measure is important when conducting counterfactuals, allowing us to evaluate how various targets change incentives throughout the entire student distribution.

With our continuous incentive measure in hand, we set out a semi-parametric approach to uncover the underlying effort response function. This is a challenging task given that effort is typically unobserved, yet variation in our incentive measure can be used to identify the effort response, as we show: Starting in a regime prior to the new accountability reform, we compare the achievement of each student against a prediction that reflects all inputs prevailing in the pre-reform period (used to construct the incentive measure). The difference between the realized and predicted test score for each level of incentive strength in this pre-reform period serves as a control for what occurs after the new incentive scheme has been implemented. Once incentives are altered post-reform, teachers and schools will re-optimize their effort choices and this will alter the corresponding outcome distribution in a systematic way, with the biggest effort responses predicted to occur where the incentives are likely to be strongest. 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 incentives.[7] By differencing the post- and pre-reform distributions, we can then uncover the underlying effort response to greater accountability.[8]

Several pieces of evidence support the view that this estimation approach uncovers an effort response. 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; and 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. Adjusting discretionary effort is an obvious channel through which performance can be altered in a manner consistent with the observed score changes; in contrast, changing education spending, varying class size and altering the assignment of teachers are more difficult to fine-tune in a short amount of time

[7] This is related to the approach in Neal and Schanzenbach (2010), who group students into deciles of the ability distribution based on a principal-component analysis using pre-reform scores, then calculate the average gain over expected scores within each decile once NCLB comes into effect. Because they only have one year of pre-reform data, they do not construct the further score difference relative to the pre-reform.

[8] We make explicit the assumptions about the technology in Section 4.

and in a way that yields the observed pattern. When we test the most plausible rival story,[9] involving schools focusing on the middle of the distribution regardless of distance from the target, we find clear evidence indicating that schools do target their effort response based on the incentive measure, according to our hypothesized channel.

Based on this reasoning, taking the estimated relationship between incentive strength and effort as given then enables us to explore the impact of different accountability schemes on the distribution of student test scores. Here, we develop a framework for counterfactual analysis in which the parameters of a given incentive scheme can be varied, in turn allowing us to compute performance frontiers that are comparable across schemes. For the main accountability systems used in practice – those setting fixed and value-added performance targets – we use the known distribution of incentive strength measures among students observed in the data for each of the two types of scheme. Varying the parameters of each scheme counterfactually while maintaining a given reference cost allows us to trace out the frontiers associated with both types of scheme in terms of their implied average performance and dispersion of scores, and further, to compare these with more refined schemes yet to be enacted. These include value-added schemes that vary the bonus depending on observed student characteristics – for instance, whether they are disadvantaged – and the maximally efficient scheme.

The types of estimate from this exercise are new to the literature. First, we show that there is a clear tradeoff between average performance and the tightness of the score distribution within the class of fixed schemes: higher performance, for a given cost, is always associated with greater dispersion of scores along the performance frontier. The same is true for value-added schemes, though the value-added frontier lies to the right of the implied frontier under fixed targeting, associated with higher effort for a fixed cost. Of note, the value-added target set in practice in North Carolina falls inside that frontier; it still raises average scores by 75 percent more than the fixed target set in practice, though leading to 20 percent greater dispersion of scores. While not typically implemented in practice, a scheme that makes bonus payments student-specific can reduce dispersion beyond that attainable under a regular value-added scheme, taking the average performance level as given.

We also show that school-level heterogeneity is important. Relative to value-added schemes, fixed schemes do especially well in schools with a greater proportion of low-performing students, while value-added schemes are vastly superior in high-SES schools.

[9] This is rival in the sense that the associated test score pattern is not explained by incentive variation.

Together, the analysis has relevance to the important issue of incentive design in education, and particularly, the refinement of existing incentive schemes. It also has broader ramifications for using policy variables to improve education productivity.

The rest of the paper is organized as follows: the next section sets out a theoretical framework that underlies our estimation approach. In Section 3, we first describe the institutional context and the rich administrative data set we have access to, followed by descriptive evidence relating to the impact of NCLB in a setting where a value-added system – the ABCs – already operated. Section 4 presents the research design. We outline our implementation of this design in a North Carolina context in Section 5, along with evidence as to its plausibility. In Section 6, we describe our framework for calculating counterfactuals, and present the findings from our counterfactual analysis. Section 7 concludes.

2 Theory

We present a simple model of the education process that links accountability incentives to school performance. This serves as motivation for our ex ante incentive strength measure, and as a means to analyze the determinants of optimal effort, which will guide our empirical implementation. We also consider the setting of accountability targets, which is part of the planner's more general incentive design problem.

2.1 Model

The model has the following basic elements:

First, there is a *test score technology* that relates measured education output y to various inputs. Give our interest in the effects of 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 actions simply as 'effort.'[10] 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.[11]

To be concrete, we write the technology as $y_i = q(e_i, \theta_i) + \epsilon_i$. We will start by focusing on performance at the individual student level, with i referring to a particular student. Output

[10]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."

[11]We will take a more narrow view in our empirical implementation, with effort referring to changes in observable test scores that are attributable to incentive variation.

y_i is that student's test score; e_i is the effort of i's instructor,[12] which is endogenous to the prevailing incentive scheme; and θ_i represents non-effort inputs such as student ability or the stock of accumulated knowledge, taken as given. Effort and exogenous student characteristics are assumed to be related in a systematic way to output, captured by the function $q(e, \theta)$; the simplest form for this would be linear. The output measure y is also influenced by a noise component, given by ϵ_i. For expositional clarity in the analysis that follows, we define $H(\cdot)$ and $h(\cdot)$ as the cumulative distribution and probability density functions of $-\epsilon_i$, functions that are common across all educators i.[13]

Second, we characterize an *incentive scheme* by a target y_i^T and a reward b, both under the control of the incentive designer. This formulation of the target allows for a range of possibilities, considered in more detail below: the target faced by the educator of i could be an exogenously fixed score; it could a function of average student characteristics, including past performance; it could 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, the educator of student i faces a convex cost of effort $c(e_i)$. We will assume the functional form of the cost function is known.

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 empirically. (In Section 4, we outline a strategy for doing so.)

Types of Scheme

To provide a sense of the model's mechanics, we now consider some common incentive schemes.

Under a *piece rate*, the 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: $b \cdot \frac{\partial q(e_i, \theta_i)}{\partial e_i} = c'(e_i)$. Of note, the latter does not include the target y_i^T, making the choice of effort invariant to it.

In practice, *threshold* schemes are far more widespread in an education setting. One reason for this is that a threshold scheme gives policymakers greater cost control – the maximum

[12] Assuming that there is just one educator for a given student is a convenient starting point. This implies that i can refer to the educator also.

[13] Defining the densities over $-\epsilon$ rather than ϵ does not change any of the theoretical predictions. The adjustment proves useful when connecting the theory to empirics, allowing us to present the results in a more intuitive way.

payout under the scheme is determinate, for example. We will focus on such schemes in what follows.

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

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

Note that, unlike the case of a piece rate, optimal effort will be a function of the target. Further, optimal effort depends on the value of θ in a systematic way. This is worth considering further, as we will be exploiting variation in exogenous determinants of performance (captured by θ) in order to construct our continuous incentive strength measure.

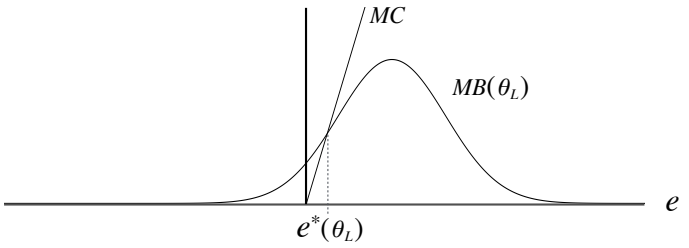
To develop intuition, take the case where $q(e, \theta)$ is simply the sum of its arguments: $q(e_i, \theta_i) = e_i + \theta_i$. The additively separable assumption implies that the marginal benefit of effort, given by the LHS of (1), simplifies to $b \cdot h[\theta_i + e_i - y_i^T]$. Holding the reward parameter b fixed, marginal benefit is then a function of two quantities. The first is the gap between the systematic component of the score, $q(e_i, \theta_i)$ and the target for school/educator i. The second is the density of the error in the performance measure, $h(\cdot)$, evaluated at that gap.

Suppose for simplicity that the error term is unimodal and also symmetrically distributed around a mean of zero. Then consider three cases, corresponding to variation in the underlying conditions governing education production in three types of school, where $\theta_L < \theta_M < \theta_H$. These could be thought of as referring to schools that educate low-, moderate- and high-ability students on average, respectively.

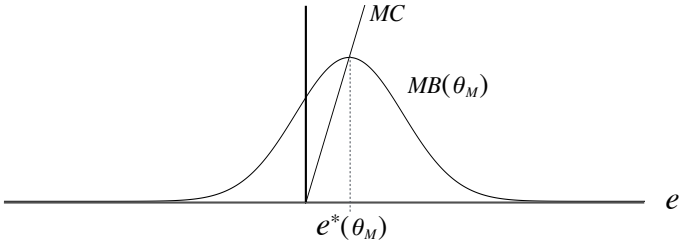
We illustrate the effects of shifting θ on optimal effort in Figure 1. Effort is on the horizontal axis, and the intersection with the vertical axis indicates zero effort – the origin for the marginal cost of effort curve in each panel. The peak of the marginal benefit curve will be found at the effort level, $\bar{e}$, for which the predicted score equals the target (and hence the argument of $h(\cdot) = 0$). 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.[14]

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

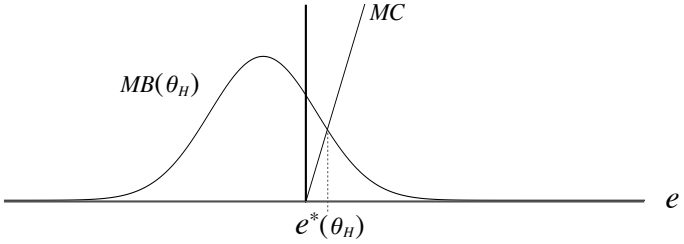
[14]Thus, the peak when $\theta = \theta_L$, in panel (a), is to the right of the peak when $\theta = \theta_M$, in panel (b), and so on.



(a) Low θ relative to y^T



(b) Intermediate θ relative to y^T



(c) High θ relative to y^T

Figure 1: Optimal Effort and Varying Exogenous Production Conditions under a Threshold Scheme

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 $\theta = \theta_H$. Here, the underlying production conditions are *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 $\theta = \theta_M$.

To summarize, several points emerge from this simple illustration that are relevant to our approach: 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. The location of the density function will be affected by exogenous conditions.

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.[15] Related to this, 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.[16]

Fourth, the optimal effort response is likely to have an inverted-U shape.[17] 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: 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 design.

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

[16] See Section 4.

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

Target setting

Having discussed the simple analytics of target threshold schemes in general, we now consider the way targets are, and can be, set. Given that schemes involving fixed and value-added targets are the most widely used in practice, we turn to those first.

Fixed schemes: These involve targets that are the same for all agents involved with a certain grade g.[18] Let the test score of student i in grade g be given by y_{ig}. The grade-specific target y_{ig}^T that applies for that student i under a fixed scheme can be written y_g^T. The fixed scheme sets a threshold: $b \cdot 1_{y_{ig} \geq y_{ig}^T}$, 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 $y_{ig}^T = \alpha_g y_{ig-1}, \ \forall \ j$ 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 (as with schemes in North and South Carolina and Florida).

Fixed and value-added schemes can be viewed as special cases of a more general class of incentive scheme. Given prior information I, these are characterized by reward $b(I)$ and target $y^T(I)$, equivalently written as $\{b(\hat{y}), y^T(\hat{y})\}$, where $\hat{y}$ is the score predicted using all prior information available to the econometrician.

To illustrate the more 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 linear 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. 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 $\{\alpha_p\}_{p=0}^{P}$. In this setup, the target under a fixed scheme imposes $\alpha_0 = \alpha$, and $\alpha_p = 0$ for all $p > 0$. 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 benchmark. Given the definitions, y^T can replicate $\hat{y}$ with some constant shift d. In partic-

[18] Fixed targets could also apply in common to all grades.

ular, 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, defined in Section 6 below, is a special case of this.

2.2 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: $\hat{y}_{ig} - y_{ig}^T$.

Incentive strength π: Because this gap plays a key role in the empirical implementation that follows, building on the model, 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. Under a fixed scheme, for example, its distribution might follow a bell shape, reflecting the underlying test score error distribution; under a value-added scheme, its distribution is likely to be tighter, 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, for a given b. It is worth emphasizing that the functional form of $e^*(\pi_{ig})$ is *unknown* and must be inferred empirically: Section 4 describes our strategy for uncovering this mapping in a semi-parametric way.

2.3 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 focuses on 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 straightforward to aggregate up to the classroom or school level, useful when exploring the classroom or school-wide effects if incentive schemes.

This extreme case does not do justice to the constraint that teachers are often under, making it difficult to individualize instruction. As an alternative, we now consider the other extreme classroom aggregation case where each teacher chooses an identical level of effort for

all students under her care. Let j refer to the educator in question, so that the production technology for student i is $y_{ij} = q(e_j, \theta_{ij}) + \epsilon_{ij}$. The average score for educator j is then $y_j \equiv \frac{1}{N_j}\sum_{i(j)}^{N_j} y_{ij} = \frac{1}{N_j}\sum_{i(j)}^{N_j} q(e_j, \theta_{ij}) + \epsilon_j$, where N_j is the number of students for which the educator is responsible.

Redefining $H(\cdot)$ and $h(\cdot)$ as the cumulative distribution and probability density functions of $-\epsilon_j$, and defining the classroom target as $y_j^T \equiv \frac{1}{N_j} y_{ij}^T$, we have the following first-order condition under a threshold-based scheme:

$$b \cdot h\Big[\frac{1}{N_j}\sum_{i(j)}^{N_j} q(e_j, \theta_{ij}) - y_j^T\Big]\frac{1}{N_j}\sum_{i(j)}^{N_j}\frac{\partial q_{ij}}{\partial e_j} = c'(e_j). \tag{2}$$

Assuming that teacher effort and non-effort inputs are additively separable, the condition becomes $b \cdot h[e_j + \theta_j - y_j^T] = c'(e_j)$, where $\theta_j \equiv \frac{1}{N_j}\sum_{i(j)}^{N_j}\theta_{ij}$. Therefore, under the additive separability assumption, the educator chooses a level of uniform effort according to the distance between the average student ability and the average target, $\theta_j - y_j^T$.[19] In addition, it is worth noting that class/school size effects do not play a role in the uniform effort choice under additive separability. This simple 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 (see Section 6).

3 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 useful 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 each school is assigned a school-grade-specific target gain, which depends on both average previous student performance and on 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-

[19] The more general nonlinear case will result in a level of uniform effort being chosen which depends on some other statistic of the $(\theta_{ij} - y_{ij}^T)$ distribution, which might be solved for numerically given a particular functional form.

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. As a necessary condition for satisfying AYP, a school must ensure that the percentage of its students in each subgroup achieving proficiency status on state tests meets the state-mandated target. If a school fails to meet each of its subgroup-specific targets, it faces an array of penalties that become more severe over time in the event of repeated failure.

3.1 Data

In addition to the institutional incentive variation, North Carolina has 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. The demographic student-level information also allows us to track any effects of the accountability regimes for students belonging to different demographic groups.

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 special education, vocational, and alternative schools. 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 and the (potentially) different incentives in schools with only one or two high-stakes grades. 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 1 provides sample summary statistics. Our main performance measures 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 and the 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

Table 1: Summary Statistics

Student-Level			
	Mean	Standard Deviation	Number of Observations
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	Standard Deviation	Number of Observations
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 special education, vocational, and alternative schools. We also exclude high schools and schools with a highest grade served that is below grade five. Student-level summary statistics are calculated over all third to eighth grade student-year observations from 1997-2005 in the 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.

NCLB, which requires that a given proportion of each of nine student subgroups exceeds a target score on standardized tests.

The longitudinal nature of the data set 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, though the largest gains occur in each case in the earlier grades. Student gain scores are, as noted, the focus of the ABCs program, which sets test score *growth* targets for schools, requiring that students demonstrate sufficient improvement as they progress through their educational careers.

The data set provides information about individual students' race, disability, limited English proficiency, and free lunch eligibility. In the aggregate, about 39 percent of students are minorities (non-white), 14 percent are learning-disabled, only 3 percent are limited English-proficient, and 42 percent are eligible for free or reduced-price lunch. Around 27 percent of students have college-educated parents.

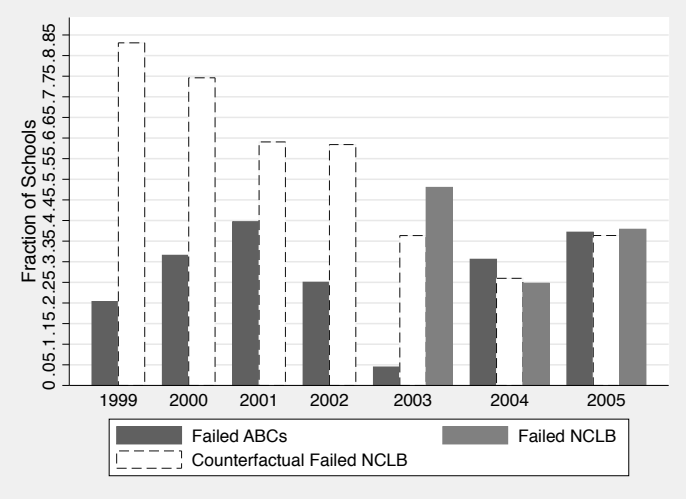
With respect to the school-level performance variables, 27 percent of schools failed the ABCs and 37 percent failed NCLB across the sample period. Recall that NCLB is a proficiency count system, which assesses school performance according to the fraction of students achieving proficiency status on End-of-Grade tests. Throughout our sample period, the average school had a school-wide proficiency rate of 79 percent on math and reading tests.[20]

3.2 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, and further, that the responses affected future incentives under the ABCs.

Figure 2 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 the years prior to 2003 by applying the NCLB outcome calculations in 2003 to the underlying

[20]The school-wide proficiency rate does not directly map into school-level NCLB outcomes, as schools are held accountable for the proficiency rates of subgroups of students in addition to the school-wide rate.



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 2: School Performance from 1999 to 2005

student-level performance data for prior years. Specifically, we first use the student-level demographic information to construct NCLB subgroup memberships prior to NCLB at each school.[21] In line with the NCLB provisions that operate in the state, we hold a school accountable for the performance of a subgroup of students only if that subgroup's membership is at least 40 students in a given year. We then use the NCLB test score and proficiency rate targets that prevailed in 2003 to determine whether all of the school's accountable subgroups satisfied AYP targets,[22] and assign the school a counterfactual NCLB 'pass' or 'fail' status in each year based on this information.

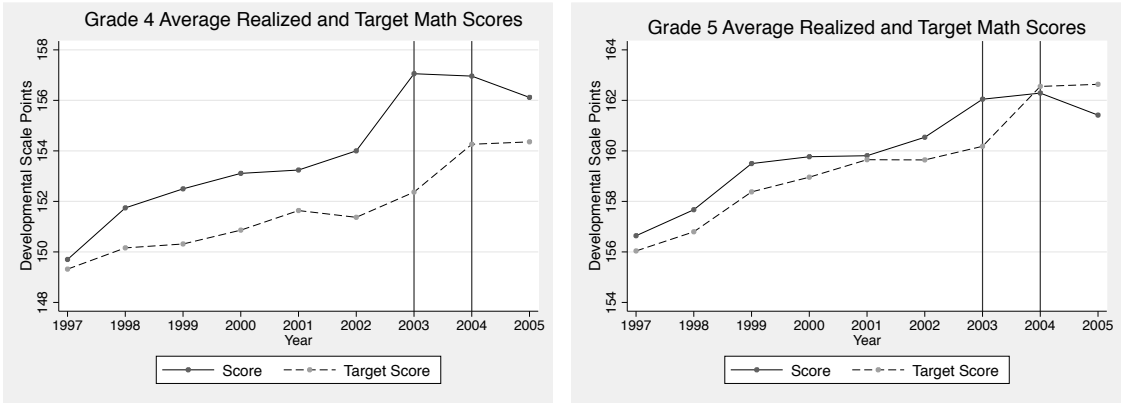
As is clear from the figure, the year in which NCLB was introduced – 2003 – is associated with a remarkable decline in the fraction of schools failing the ABCs, down from 25 percent in 2002 to only 4 percent in 2003. To lend credence to the notion that this reflects an NCLB-triggered response, notice 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.[23] The figure also shows that the ABCs improvement was short-lived, as the failure rate more than jumps back, to over

[21]The nine groups are white, black, Hispanic, Asian, Indian, multi-racial, disabled, free or reduced price lunch-eligible, and limited English proficient.

[22]For a subgroup to meet AYP targets in 2003, 74.6% of its students had to reach proficiency status in math and 68.9% had to reach proficiency status in reading

[23]Due to the many nuances associated with the implementation of NCLB over the years, 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. (Other researchers using the NCERDC data have faced similar difficulties (see, for example, Ahn and Vigdor, 2013).) In 2003, we understate the failure rate by around 10 percentage points. In 2004 and 2005, we are much closer, within a couple of points.

30 percent in 2004.



(a) (b)

Notes: Panel (a) shows the evolution of average math scores and average math targets for fourth grade, while panel (b) shows the analogous evolution for fifth grade.

Figure 3: ABCs and NCLB Interaction

Figure 3 sheds light on the way that NCLB's introduction disrupted schools' ABCs incentives and, ultimately, their ABCs outcomes. The figure reports average test scores and average ABCs score targets for math. The vertical lines depict the year in which NCLB was implemented (2003) and the year immediately after (2004). The figure shows that, relative to other years, average math scores in both fourth and fifth grade display impressive improvements from 2002 to 2003, the solid lines rising quite sharply. These improvements led to the large decline in the fraction of schools failing the ABCs shown in Figure 2.

Observing the trends around 2004, however, it is apparent that schools responded aggressively to NCLB, with the 2003 improvement making it more challenging for them to reach their ABCs targets in subsequent periods: by raising test scores in response to NCLB, schools improved contemporaneous ABCs performance but simultaneously increased future ABCs targets and, correspondingly, lowered the likelihood of future ABCs success (as indicated by the steeply rising dashed lines in both panels). The dynamic interaction between the two schemes is seen clearly by observing the pronounced increase in the average fifth grade target in 2004 – a direct consequence of the pronounced increase in the average fourth grade score in 2003. In fact, 2004 marks the first year in which the average fifth grade target rose above average performance, implying that the *average school* failed to satisfy its fifth grade math requirement. Thus, by improving students' performance in 2003 so drastically, many schools set baseline scores that they were unable to improve upon enough in 2004, thereby incurring the real cost of missing out on their performance bonuses.

4 Research Design

The research design presented in this section is central to our analysis. Our goal is to uncover the optimal effort response, $e^*(\pi)$, described in Section 2 for a given incentive strength π. Building on the descriptive evidence in the previous section, the strategy we follow makes use of the introduction of the new performance requirements under NCLB as an *exogenous* shock to the school decision process occurring in 2003, the first year of the reform. 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 uniform, irrespective of incentive strength π.[24] 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 π – our *ex ante* incentive measure – as we are interested in seeing how changes in formal incentives are reflected in the score distribution in a manner attributable to an effort response.

To that end, consider a reform R that introduces new performance targets $y_R^T(\hat{y}_R)$ for educators,[25] thereby changing the incentives to exert effort. We will write scores in this post-reform environment, using the linearity assumption, as $y_R(\pi) = \hat{y}_R + e^*(\pi) + \epsilon_R(\pi)$, expressed as a function of π.

Incentive Strength

To construct our ex ante incentive strength measure, we distinguish 2003 – the year in which the new incentives came into effect – from earlier years. We first flexibly predict student

[24] Such uniformity can be checked, to some degree, by plotting the pre-reform test score distribution, intuitively to see whether there are any 'bumps.' We provide descriptive evidence in the next section.

[25] Note that $\hat{y}_R$ represents the predicted score in the post-reform environment, excluding any additional effort response, e^*, to the reform.

performance in those earlier years using several covariates, including lagged test scores.[26] We then predict performance in 2003 using student covariates from that year (including lagged scores) along with the saved coefficients from the prediction exercise in pre-reform period.

Our *ex ante* incentive measure is then defined as the difference between the predicted value (which does not include *additional* effort in 2003) and the target; in terms of the model above, this value is denoted as $\pi \equiv \hat{y} - y^T$. Given the way we construct it, the predicted score component is invariant to any changes occurring in 2003 or in later years. 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.

Effort Response

With the *ex ante* incentive strength measure in hand, we then turn to the main task: to determine the optimal effort response for each value of our continuous incentive strength measure, π. We do this in the following way: for each value of π, we compute the difference between the realized and predicted scores. We do so for years 2000 to 2002 (serving as a control for preexisting patterns) and for 2003, the year when the incentive shock occurred. Intuitively, while the former contains noise only, the latter contains noise *and* effort. In terms of the above technology, in the pre-reform period, $y(\pi) - \hat{y} = \epsilon(\pi)$, while post-reform, $y_R(\pi) - \hat{y}_R = e^*(\pi) + \epsilon_R(\pi)$.[27]

We then take the further difference between the 2003 and 2000 to 2002 distributions to isolate 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. 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. We consider supportive evidence next.

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

[27] To avoid any confusion, the pre-reform predicted score $\hat{y}$ is calculated using pre-reform prior inputs and the coefficients relating those inputs and outputs, estimated from the pre-reform period. The post-reform predicted score $\hat{y}_R$ is calculated using post-reform prior inputs along with the same pre-reform coefficients.

5 Results

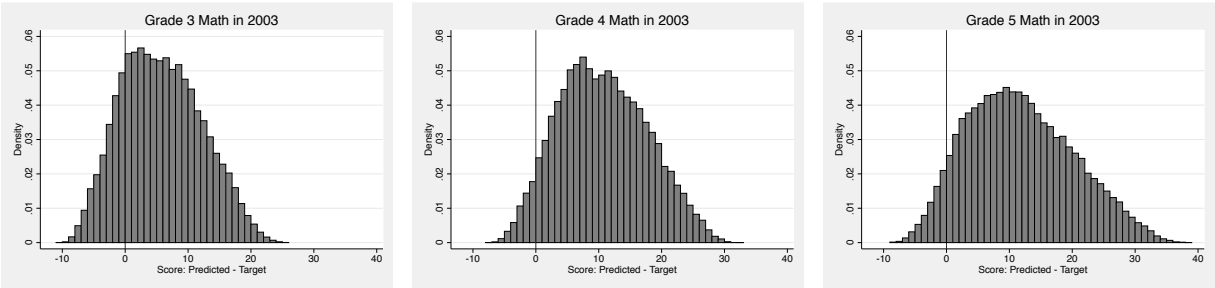
In this section, we show results from the implementation of our research design, along with evidence relating to the validity of the approach.

5.1 Testing for Bunching

We start with 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.

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 exogenous.



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

Figure 4: Distribution of Predicted Scores minus the NCLB Target

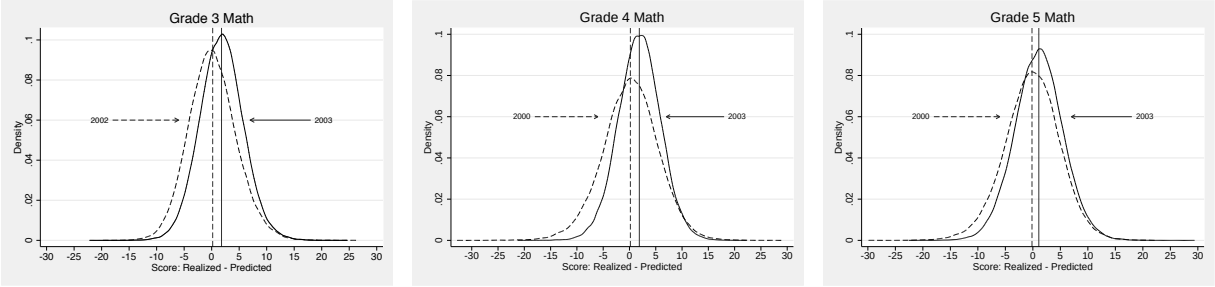
5.2 The Test Score Response

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

Figure 5 shows the densities of realized minus predicted test scores in both the pre-period (2000 and 2002)[28] and the post-period (2003), which we interpret as the densities

[28]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

of unobservable test score determinants, including the effort of educators. Predicted scores



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

Figure 5: 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 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 2. These schemes reward schools (or refrain from punishing them) for 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 who are 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 6 shows that these are exactly the patterns we find across the predicted test score distribution. In 2003, the gains above predicted scores are low for students who are predicted to be far below the proficiency threshold; they begin to increase for students who

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.

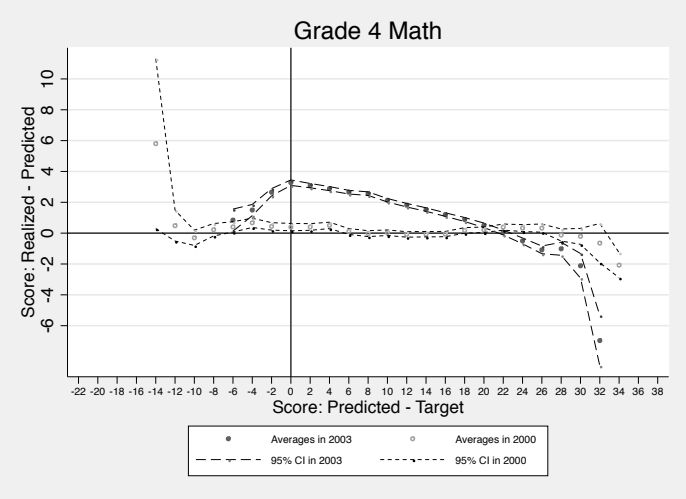


Figure 6: NCLB Effort Response

are predicted to be close to the threshold; and they decline again for students who are predicted to be far above the threshold. Of note, 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.

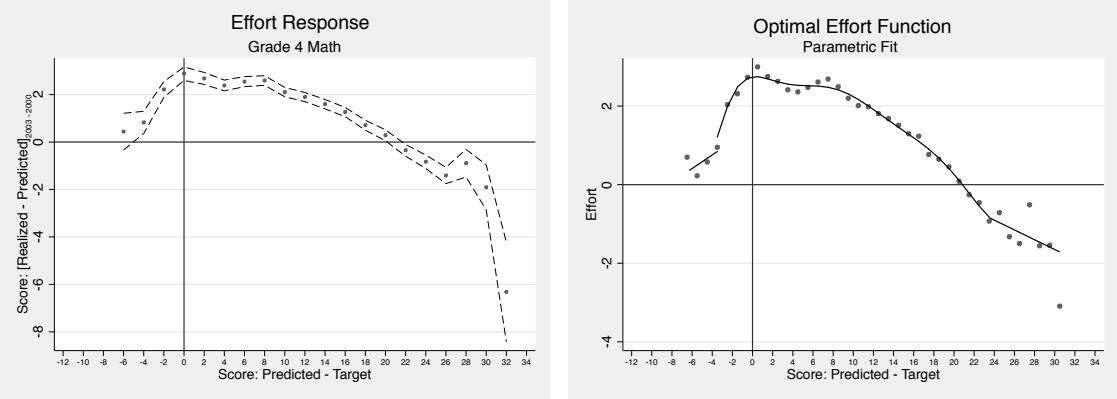
Generally speaking, to the extent that educators care about incentives under the new regime, adjusting discretionary effort is an obvious candidate input through which performance can be altered, and in a manner consistent with the observed change in test score profile. In contrast, changing education spending or altering the assignment of teachers, or other possible channels do not seem likely candidates to explain the observed changes. We therefore take the evidence to support the view that teachers redirected their efforts in response to NCLB, and in a manner consistent with the optimal response to a proficiency-count system.

5.3 Incentive Strength and the Optimal Effort Function

In Figure 7(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. In Figure 7(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 7(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 a standard production technology assumption used in the literature.

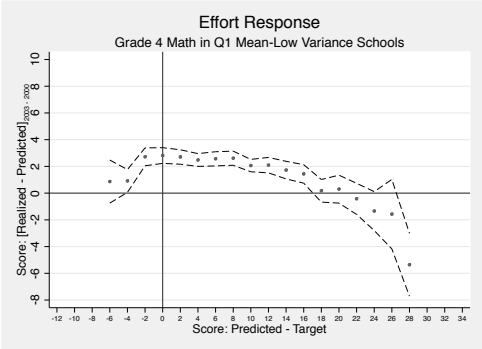


(a) 2003 minus 2000 (b) Optimal Effort Function

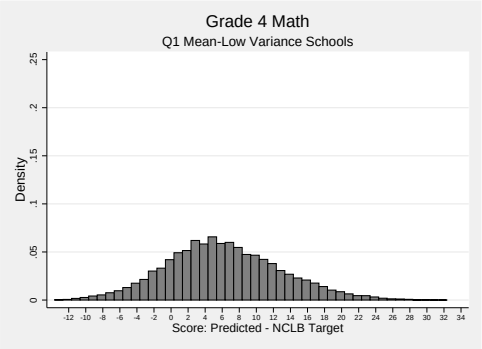
Figure 7: Derivation of the Optimal Effort Function

Rival Effort Story

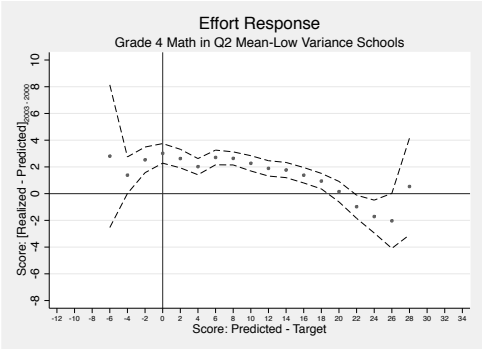
Our maintained hypothesis is that we are uncovering the effort response with respect to the incentive strength measure, π. As an alternative, effort might vary with respect to a student's relative position in the predicted score ($\hat{y}$) distribution within his or her school. For example, it is possible 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



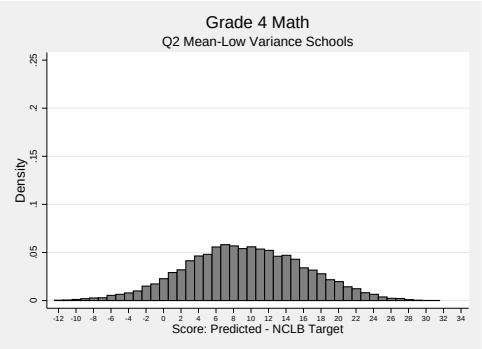
(a) Effort in Q1 Mean Schools



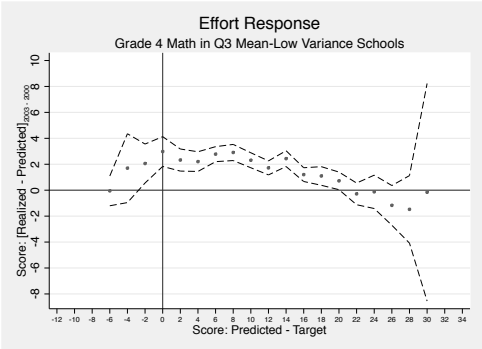
(b) π Density in Q1 Mean Schools



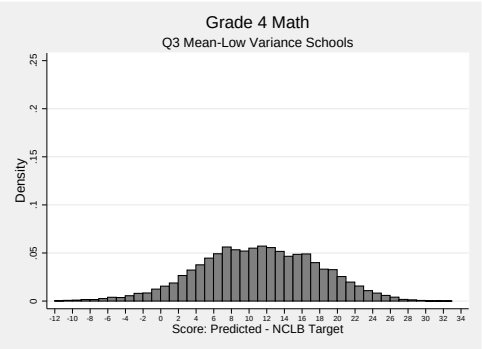
(c) Effort in Q2 Mean Schools



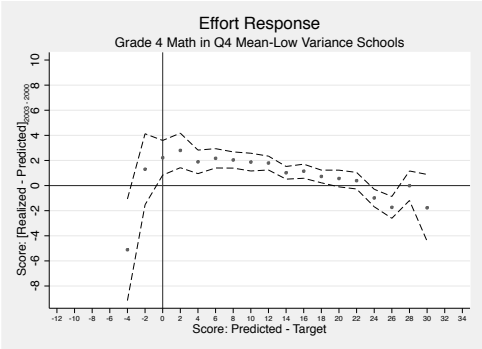
(d) π Density in Q2 Mean Schools



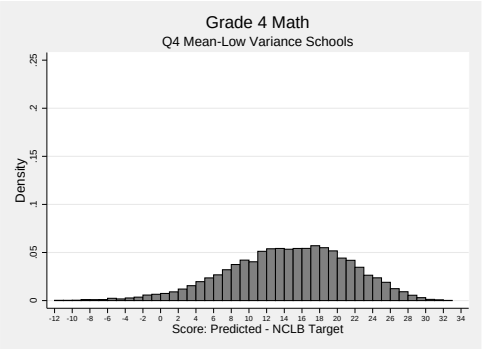
(e) Effort in Q3 Mean Schools



(f) π Density in Q3 Mean Schools



(g) Effort in Q4 Mean Schools



(h) π Density in Q4 Mean Schools

Figure 8: Responding to π not the Relative Position of $\hat{y}$?

in the ability distribution where π equalled zero, then varying π counterfactually to make inferences about competing accountability schemes would seem unwarranted.

To assess this possibility, we 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 8 plots the effort responses and π densities separately for schools in each of the quartiles of the mean, focusing on below-median variance.[29] 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.

6 Counterfactual Simulations

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

6.1 A Framework for Counterfactual Comparisons

We develop a framework that enables us to compute the implied test score distribution for a particular set of targets $\{y_{ig}^T\}$, given two key elements: the optimal effort function (already discussed), and the implied distribution of incentive strength measures, which we describe below. The framework allows the test score distribution associated with a particular scheme to be recovered, in turn providing the basis for calculations involving useful summary measures – the score distribution's first and second moments, for example. Further, we can trace out performance frontiers associated with a given type of scheme, and also compare

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

frontiers across schemes.

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 faced by educators. Computationally, this requires calculating $\hat{y}_{ig} - y_{ig}^T$ for each educator i in grade g, and storing the entire set of distances – in Figure 4, we showed density plots for given grades and given target schemes.

A scheme R will imply a determinate target for each educator. In terms of each educator's *predicted* score, we assume – looking ahead to our actual implementation – that the predicted scores are 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 (and in turn, the actual score, assuming effort changed), it will not alter the predicted score that feeds into the calculation of the distribution of incentive strength.[30]

Formally, we will write the density of π under scheme R as $f_R(\pi)$, on the support $\pi \in [\underline{\pi}_R, \bar{\pi}_R]$. Our approach will involve recovering this density semi-parametrically.

Counterfactual Outputs

As a starting point, we focus on two appealing summary measures: average effort associated with a given scheme and the corresponding variance of scores.

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

For comparability, we will use the average cost 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 $\tilde{\pi}_{ig} \equiv \pi_{ig} + \epsilon_{ig} = y_{ig} - y_{ig}^T$, 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.[31]

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

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

One of our preferred summary measures, *average effort*, of scheme R with y_R^T *relative to* scheme R' with $y_{R'}^T$ can be calculated using Ω_R, $\Omega_{R'}$, C_R and $C_{R'}$, defined above. One might wish to compare average effort under fixed versus value-added targeting – i.e. $\Omega_{R=F}$ versus $\Omega_{R'=VA}$, for instance. This cannot be done directly, however, if $C_{VA} \neq C_F$. Our solution is to adjust b under the value-added scheme until $C_{VA}(b') = C_F(b)$, and then compare $\Omega_F(b)$ to $\Omega_{VA}(b')$.[32] We will use the measure, $\frac{\Omega_F(b)}{\Omega_{VA}(b'),}$ to compare performance under these alternative schemes.

Now turning to the dispersion of scores under scheme R, define $\tilde{y} \equiv \hat{y} + \epsilon$ – essentially, the test score absent any effort component. The variance in test scores is then:

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

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. 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.

Alternative Types of Scheme

Following this discussion, we can now revisit *uniform schemes*, first referenced in Section 2 above. Under such schemes, recall that the target $y^T = \hat{y} - d$, where d is some constant shift. Then 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 $C(d) = b \int \mathbf{1}_{d+e^*(d)+\epsilon \geq 0} h_\epsilon(\epsilon) d\epsilon$.[33] 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 will explore in our counterfactual analysis next. Generally, a heterogeneous reward has implications for both average effort and the dispersion of scores, but it only affects the dispersion under a uniform scheme. Here, increasing the weight on students with the lowest $\hat{y}$ will serve to lower the spread. If $e^{max} \ll \sigma_{\hat{y}}^2$, then there will be a positive lower bound for the variance.

Using a heterogeneous reward to reach that minimum variance under the maximally ef-

[32] If e^* peaks near $\pi = 0$, then average effort will increase under a value-added scheme, since a greater proportion of students are marginal in that case.

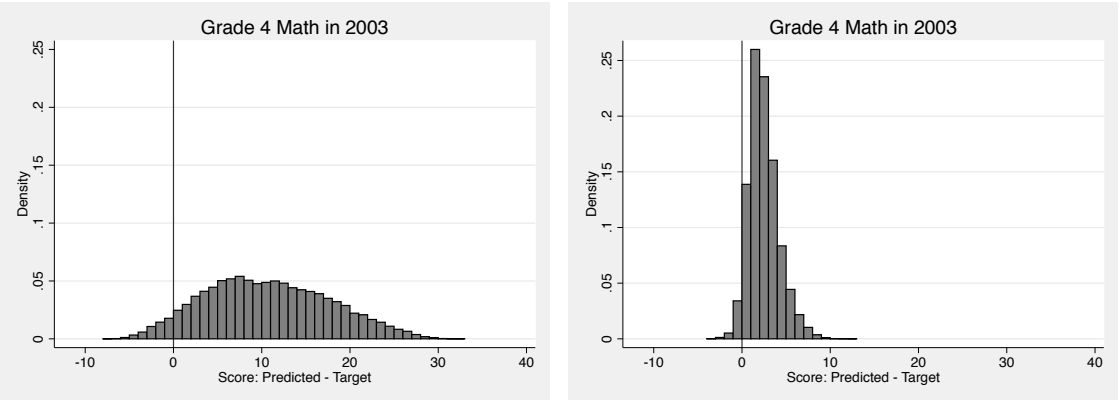
[33] In Section 2, we define $h(\cdot)$ as the probability density function of $-\epsilon$. Here, we use $h_\epsilon(\cdot)$ to denote the probability density function of ϵ.

ficient scheme is socially optimal only if it is sufficiently transparent to agents; otherwise, the effort response may be substantially attenuated in comparison to simpler schemes. Yet it still serves as a useful yardstick in the comparisons below.

6.2 Baseline Counterfactual Results

In describing the results from our counterfactuals, we first a set of fixed proficiency targets and a set of value-added ('VA') proficiency targets, deriving the implied distributions for π in each case. For the set of fixed proficiency targets, we simply move the proficiency threshold below and above the NCLB threshold: for the VA proficiency targets, we vary the multiplicative coefficient on the prior math scores in the ABCs math target, moving it below and above the coefficient that prevails under the ABCs. 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.

As an example of the proficiency targets we consider, Figure 9 shows the distributions of π that prevail under the NCLB proficiency target and the ABCs VA target. 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 proficiency targets will be much more uniform across the distribution of students than those under fixed proficiency targets.



(a) NCLB Target (b) ABCs Target

Figure 9: π Densities

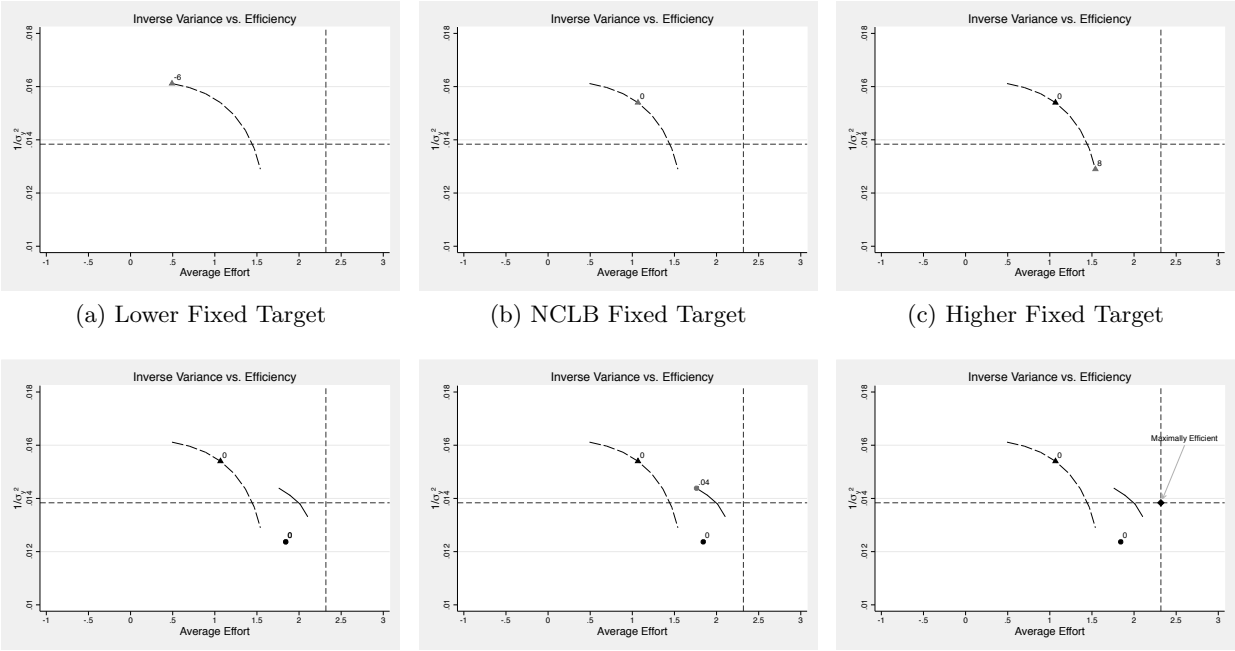
We present the results of selected counterfactual simulations in Figure 10. To determine the test scores that prevail under counterfactual scheme R, we define test score outcomes as

the sum of predicted test scores, student-specific effort, and normally-distributed noise:

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

For each simulation, we obtain two summary measures of interest: (1) average effort across students and (2) 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, shown in Figure 10f) is defined as average effort when all students are assigned the common level of π that maximizes the ratio of average effort to average cost. For each counterfactual regime, we equate the cost to the cost prevailing under the 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 10 represent inverse variance and average effort 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 10: Counterfactual Simulations

There is a clear tradeoff between variance and average effort. To explain the tradeoff for

the class of fixed targets, consider the target that is 6 points below the NCLB target in Figure 10a). 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 substantially lower average effort. 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 efficiency and inequality. 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. Evidently, 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 10d), 10e), and 10f) 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. We explain this result as follows: First, as shown in Figure 9b), the π distribution under the ABCs VA target is 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. Second, 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. 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).

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, 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 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 10 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.

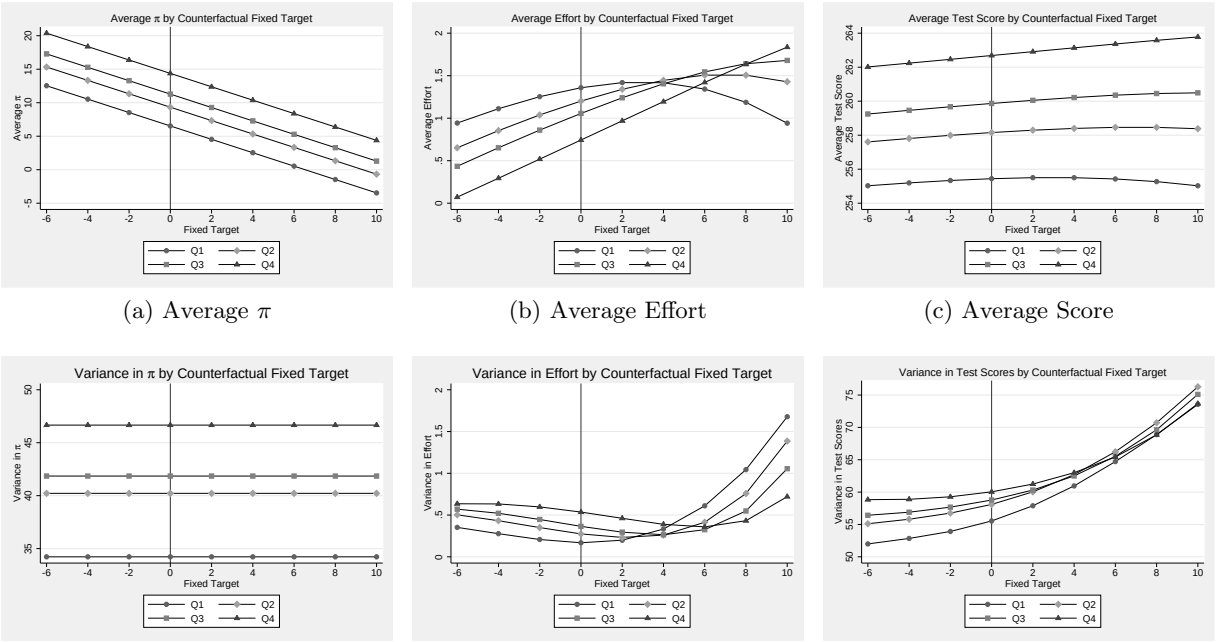
Turning to a comparison of specific targets, the NCLB fixed target balances average effort and inequality with respect to other fixed targets. It results in 117 percent more effort and 4 percent higher variance than the lowest fixed target, and it produces 31 percent less effort and 16 percent less variance than 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 preexisting 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.

6.3 Exploring Heterogeneity Across Schools

6.3.1 Aggregate Differences by School Type

Having documented clear tradeoffs between average effort and the variance in student test scores, in this section we investigate how counterfactually manipulating the proficiency targets affects the distributions of π, effort, and realized scores, both within and across schools.

To get a sense of how these objects evolve as targets change, Figure 11 groups schools by quartiles of the school-specific mean value of π that prevails under NCLB and then plots changes in the mean and variance of π, effort, and realized scores at each type of school. The horizontal axes measure the distance of the counterfactual fixed target from the NCLB target.



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

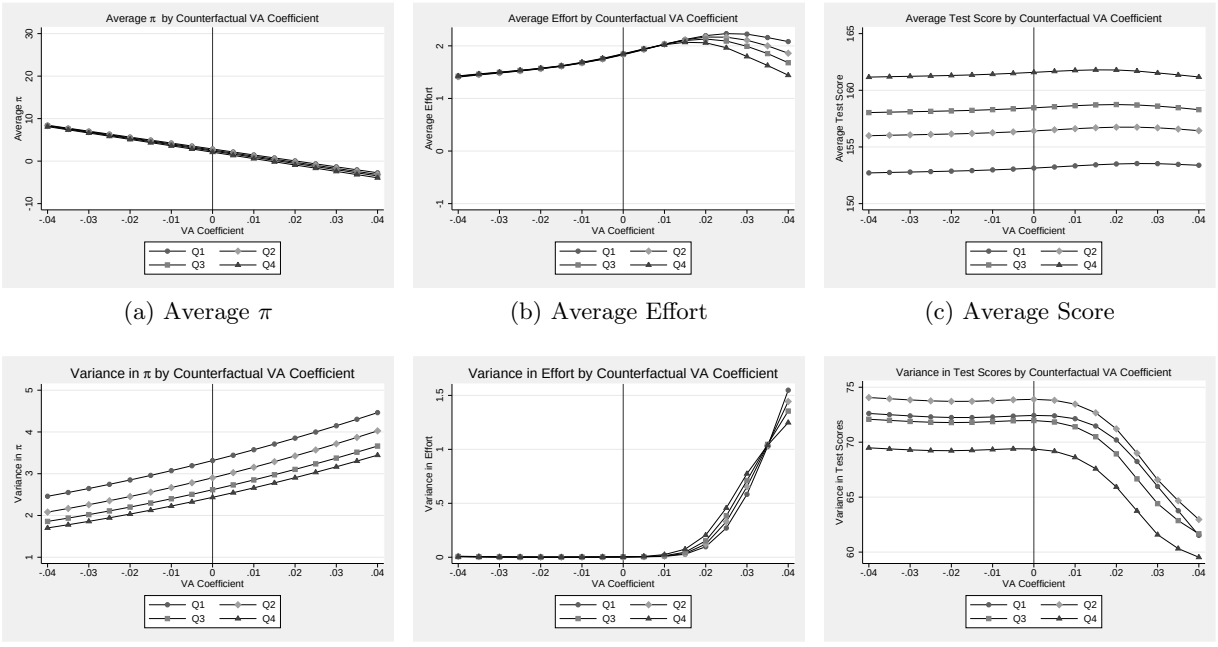
(d) Variance of π (e) Variance of Effort (f) Variance of Score

Figure 11: Counterfactual Fixed Targets and School-Type Heterogeneity with Student-Specific Effort

Panels (a) and (d) 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 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 rapidly increase at higher values of the target. The target corresponding to the minimum variance is lower at schools in lower

quartiles, as low targets achieve a relatively symmetric (around the target) π distribution at low-performing schools while high targets do the same at high-performing schools. At the highest fixed target, the variance of effort at quartile-one schools is about double the variance of effort at quartile-four schools.

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.[34]



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

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

Figure 12: Counterfactual VA Targets and School-Type Heterogeneity with Student-Specific Effort

Figure 12 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).[35] A key insight from Figure 12 is that there is much less (in fact, almost

[34]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.

[35] Note that, because North Carolina's ABCs program uses coefficients and targets exclusively with test scores

zero) heterogeneity in the average value of π across school types. VA schemes effectively set student-specific targets, implying that similar incentive strength is attached to each student, regardless of his or her preparedness. 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 in class of 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.

The effort function is relatively flat for most values of π and then quickly declines once π falls below -2. For values of the multiplicative coefficient ranging between -0.04 and 0.01, the π distribution has a relatively small variance and is in the domain of the effort function where the corresponding values of effort are fairly uniform. 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 suddenly increase the dispersion in effort across students at each type of school. In panel (f), one can see these dynamics in 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.

6.3.2 Within- and Between-Schools Variances

In this section, we document the implications of the dynamics discussed above on 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 2 decomposes the variance of π, effort, and realized scores into the (a) the variance within schools, (b) the variance between schools, and (c) the amount of the within-schools variance that occurs within classrooms.

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.

Table 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.

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 to 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-schools variance in test scores is 33 percent higher under the fixed target that is ten points above NCLB than the target that is six points below NCLB; the between-schools variance is 53 percent higher.

While higher fixed targets increase the variance of tests scores, higher VA coefficients decrease it. Panel (B) of Table 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. This may be an artifact of the assumption that teachers can choose to flexibly adjust their effort across students on a student-specific basis. In reality, it is likely the case that the effort each student receives is some combination of a student-specific component his or her teacher chooses and a classroom-specific component received by all students who share the same teacher. Our counterfactual simulations thus far effectively put zero weight on the classroom-specific component, instead allowing teachers to perfectly discriminate across all students in their classes.

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 this 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 section reports results under the assumptions that each student in a classroom receives the same amount of effort and that teachers reach decisions over classroom effort levels according to the model in Section 2.3.

6.3.3 Classroom-Specific Effort Constraint

Table 3 shows how the variances of each variable evolve when effort is constrained to be equal within classrooms. As in Section 2.3, 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 strenght, π, the variances of π are the same as those in Table 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.[36]

Table 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.

[36] 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 35.

Under the common effort assumption, the between-schools 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, with the within-classroom variance explaining less of the within-school variance in test scores for higher fixed targets and more for higher VA coefficients.

While neither likely represents a completely accurate depiction of teachers' real effort choices, the separate consideration of the two extreme cases – student- versus classroom-specific effort – helps us reasonably bound the effects on the distribution of outcomes from counterfactually varying the proficiency target.

6.3.4 Differential Effort Functions Across School Types

While teachers are likely constrained to some degree by how much 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. 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. Essentially, 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 (un)likely to pass no matter the actions it takes, it may not respond to the accountability provisions at all.

We explore this type of of heterogeneity to some extent in Figure 8 in Section 5.3, 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 section by recalculating the optimal effort function within each school

type and redoing the counterfactual analyses using these school-type-specifc 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.

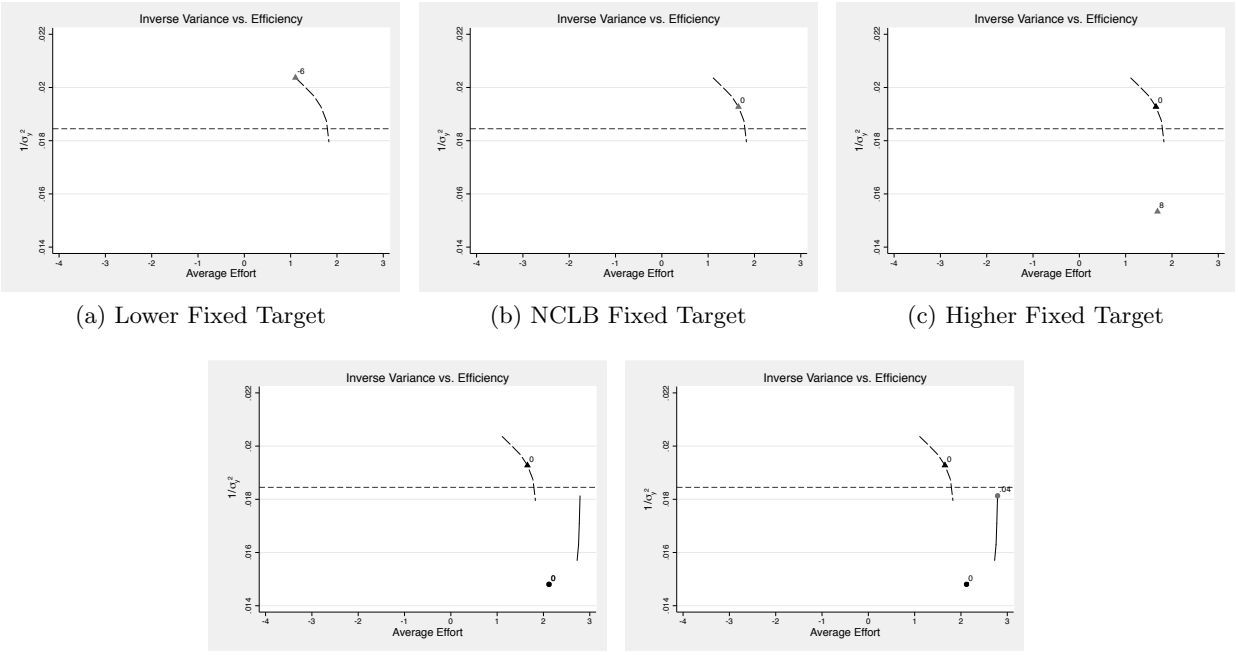
Low-Mean, Low-Variance Schools

Figure 13 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 CLB.

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 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 may 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 these 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.



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

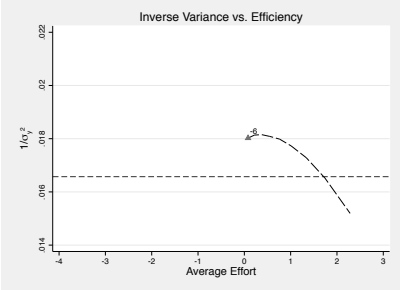
(d) ABCs VA Target (e) Higher VA Target

Figure 13: Counterfactual Simulations for Low Mean, Low Variance Schools

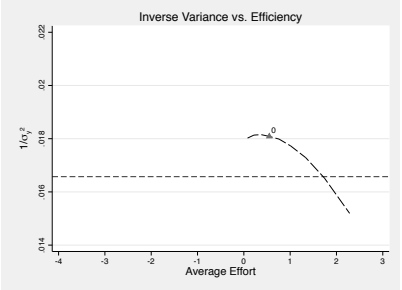
High-Mean, Low-Variance Schools

Figure 14 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.

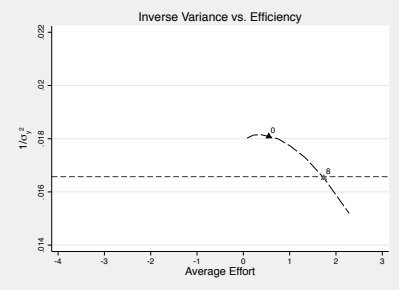
Since high performers represent a relatively high fraction of students in these schools, teachers do not redirect resources away from them to the same degree that they do in low-performing schools, which results in an effort function with small negative values of effort for large values of π and a relatively 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.



(a) Lower Fixed Target



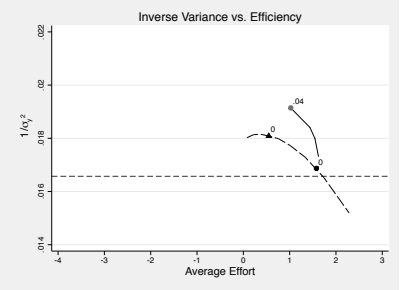
(b) NCLB Fixed Target



(c) Higher Fixed Target



(d) ABCs VA Target



(e) Higher VA Target

Figure 14: Counterfactual Simulations for High Mean, Low Variance Schools

7 Conclusion

In this paper, we have provided a new framework for measuring the performance of different incentive schemes in education on a comparable basis, useful for addressing policy design issues. After using exogenous variation in incentives semi-parametrically to identify the effort response of North Carolina teachers and schools to a prominent accountability reform, we computed the average effort and dispersion of scores associated with rival schemes, tracing out corresponding frontiers in performance space.

Among the main findings, our estimates make clear the tradeoff between average effort and a measure of spread – the inverse variance of scores – for fixed target schemes. A similar tradeoff arises for value-added targets. Value-added targets with heterogeneous bonus payments across students dominate most fixed targets in terms of the effort-inverse variance tradeoff. And we show that school heterogeneity is important, the evidence suggesting that fixed schemes perform better in schools with a greater proportion of low-performing students.

In related work, we explore dynamics in an interacted incentives environment, motivated by the descriptive evidence shown above. We showed how schools' significant outperformance in 2003 resulted in substantially higher future ABCs targets in 2004 and 2005, jeopardizing future rewards, our aim being to examine the propagation of efficiency and inequality through

this path dependence.

We are also examining how the exogenous variation we have uncovered can be used to shed light on the nature of the underlying production technology in education. Building on the analysis in this paper, we have a strategy for uncovering unobservable effort, which can be fed into a flexibly-specified education production function, allowing for nonlinearities and differential persistence of inputs. Such estimates should be potentially very relevant for policy. Our research in this area is ongoing.

References

Ahn, Tom and Jacob Vigdor. 2013. "The Impact of NCLB's Accountability Sanctions on School Performance: Regression Discontinuity Evidence from North Carolina." *University of Kentucky Working Paper.*

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.

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." Working Paper 19444. NBER, Cambridge, MA.

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." Working Paper 8855. NBER, Cambridge, MA.

Jacob, Brian A. and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics*, 118(3): 843-77.

Jacob, Brian A., Lars Lefgren, and David P. Sims. 2010. "The Persistence of Teacher-Induced Learning." *Journal of Human Resources* 45(5): 915-943.

Ladd, Helen F. and Douglas L. Lauen. 2010. "Status versus Growth: The Distributional Effects of School Accountability Policies." *Journal of Policy Analysis and Management*. 29(3): 426-450.

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

Macartney, Hugh. 2014. "The Dynamic Effects of Educational Accountability." NBER Working Paper 19915. NBER, Cambridge, MA. Forthcoming in the *Journal of Labor Economics*.

Muralidharan, Karthik and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy*. 119(1): 39-77.

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." Working Paper 16745. NBER, Cambridge, MA.

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