A Unifying Framework for Education Policy Analysis*

Hugh Macartney† Robert McMillan‡ Uros Petronijevic§

November 6, 2016

*We would like to thank Raj Chetty, Damon Clark, John Friedman, Caroline Hoxby, Juan Carlos Suarez Serrato, and seminar participants at Chicago Harris, Columbia, Duke, McMaster University, the University of Ottawa and the NBER Public Economics meeting for helpful comments and suggestions. Mike Gilraine (University of Toronto) provided outstanding research assistance. Financial support from SSHRC and 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, York University, Vari Hall, 4700 Keele Street, Toronto, Ontario M3J 1P3, Canada. Email: upetroni@yorku.ca

Abstract

This paper develops a framework that allows us, for the first time, to compare the cost effectiveness of widely-discussed education policies, including incentive-based reforms. Central to the cost-effectiveness calculation, we propose an empirical strategy to separate the overall impact of teachers on student achievement into an incentive-varying component – labeled teacher *effort* – and incentive-invariant teacher *ability* while imposing minimal assumptions. The strategy draws on exogenous variation in the incentive strength of a well-known federal accountability scheme and rich administrative data covering all public school students in North Carolina over time. Our results from identifying contemporaneous teacher effort and ability separately indicate that a one standard deviation increase in ability is equivalent to 18 percent of a standard deviation increase in student test scores, versus a 5 percent increase in scores for a one standard deviation increase in effort. Further, the overall effect of teachers (measured by value-added) is increasing in accountability incentive strength. We then implement a structural estimation procedure to uncover the persistence of ability and effort, showing that effort affects future scores positively, though by less than ability. Combining these estimates with a model of school decision-making allows us to place incentive-based reforms on a common footing with a range of alternative education policies. For illustration, we show that incentive-based reforms can be more cost-effective than reforms that target teacher ability as a means of improving average student performance.

Keywords: Education Policy Analysis, Incentive-based Reform, Cost Effectiveness, Teacher Value-added, Incentives, Accountability, Education Production, Teacher Effort, Teacher Ability, Persistence

I. Introduction

Education policy remains high up the public policy agenda, not only because of its potential for building fundamental skills in society, but also because of a pervasive sense that many public schools continue to fail. The search for viable policies to improve public school outcomes is reflected in two strands of recent education research. One influential strand assesses the importance of teachers in the production of student achievement, estimating sophisticated value-added ('VA') measures that seek to capture the overall performance impact of a given teacher (see, for example, foundational papers by Chetty, Friedman and Rockoff 2014a,b.) These types of measure have become the focus of policy interventions that include, controversially, dismissing low value-added teachers (Hanushek 2009). A second prominent strand of work studies the impact of accountability incentives on student and school performance – an issue not taken up in the value-added literature. A number of persuasive papers show that accountability schemes, which have become widespread in the United States over the past two decades, have succeeded in improving student achievement in a variety of settings.[1]

Given these separate literatures, it is natural for policy makers to wonder how policies that target teacher quality and policies that provide explicit incentives for teachers and schools might compare. This calls for a way of computing the benefits of each in terms of student achievement against the costs of the respective reforms. Yet the existing literature does not offer a means for placing the two types of policy on a common footing.

This paper sets out a unifying framework to allow such a comparison for the first time. Our starting point is to draw a formal distinction between two inputs into education production: teacher effects that are invariant to the incentive environment and those that are responsive to it; for convenience, we label the former 'teacher ability' and the latter, 'teacher effort.'[2] This distinction in hand, we then provide a general approach for identifying the effects of teacher ability and teacher effort in terms of student scores, invoking minimal as-

[1] See, for instance, Carnoy and Loeb (2002), Lavy (2002, 2009), Hanushek and Raymond (2005), Dee and Jacob (2011), Muralidharan and Sundararaman (2011), and Imberman and Lovenheim (2015).

[2] These labels are not entirely in line with common parlance. We will take effort to be any incentive-related action that raises scores, including (but not limited to) devoting more time to lesson planning, interacting more with students' parents, and organizing extra tutoring sessions after school and on weekends. While ability is sometimes viewed as being amenable to change (perhaps through teacher training), our measure of ability will equal the component of teacher value-added that does not change over time, conditioning on teacher experience.

sumptions. The approach speaks to the benefit side of the policy calculation – specifically, the benefits of permanent teacher 'ability' (and policies that influence the composition of the teacher pool) versus teacher and school effects that can be altered by incentive policies.

Our strategy exploits exogenous incentive strength variation associated with the federal No Child Left Behind Act of 2001 ('NCLB') – a proficiency system that sets achievement targets along with penalties for target non-attainment – and rich administrative data covering all public school students in North Carolina over time. As is well-appreciated in the literature (see Reback 2008, for instance), proficiency schemes like NCLB make students matter differentially at the margin according to their likely test performance relative to a fixed performance target. Drawing on this insight, our approach compares teachers who teach classrooms with higher versus lower fractions of marginal students, both before and after the incentive reform was implemented in North Carolina in 2003.

Since we observe many teachers before and after, we are able to estimate teachers' incentive-invariant ability levels using standard VA methods in the pre-reform period. Teacher-effect estimates of this type have been shown to be unbiased predictors of teachers' average impact on student test scores and important long-run outcomes (see Kane, McCaffrey, Miller, and Staiger 2013, Chetty *et al.* 2014a, and Kane and Staiger 2014). Once NCLB is implemented, however, we demonstrate that these performance measures vary systematically with incentives, while holding constant estimated teacher ability from the pre-reform period. Accounting for the incentive shock's impact on effort in a semi-parametric way allows us to separate out the relative importance of teacher ability and effort contemporaneously: a one standard deviation increase in teacher ability is equivalent to 18 percent of a standard deviation increase in student test scores, while an analogous change in teacher effort accounts for 5 percent of such an increase.

We then use measures of teacher ability and effort to investigate the extent to which each input persists in determining future test scores. To estimate the persistence of teacher ability, we rely on data from the pre-NCLB period and methods used previously in the literature, in which we regress future test scores on standard VA control variables and our measures of teacher incentive-invariant ability. In line with prior work (Jacob, Lefgren, and Sims, 2010; Chetty *et al.* 2014a), we find that approximately 40 percent of the initial effect of teacher ability persists after one year and that 20 percent remains after four years.

Estimating the persistence rate of teacher *effort* is more challenging. The ideal experiment for cleanly identifying the persistent effects of the NCLB-related effort shock in 2003 would amount to repealing both NCLB and the state's accountability program – the ABCs of Public Education – in 2004 and beyond. In that case, no future effort decisions would be made, allowing us to estimate how effort levels from 2003 (estimated in the first part of our empirical analysis) persisted to affect future test scores, based on regressions analogous to those used to estimate the persistence of teacher ability.

In practice, both accountability schemes continued to operate, and as a consequence, educators faced two contemporaneous effort decisions every period after NCLB's introduction, each affecting contemporaneous test scores while also being correlated with the initial effort response in 2003. Given these features of the observational setting, our approach involves controlling for the contemporaneous test score effects stemming from these two ongoing programs while estimating the persistence of effort from 2003. We do so using our model to develop a structural estimation approach – one that also accounts for the fact that contemporaneous NCLB effort is a function of the persistence effect that we are trying to estimate.[3]

Structural estimates of the persistence of effort reveal that 13 percent of the initial effort effect persists one year ahead, which amounts to approximately 32 percent of the one-year persistence of teacher ability. Not accounting for the test-score effects of contemporaneous effort decisions results in an overestimate of the persistence rate of effort, with a much higher fraction – 50 percent – of the initial effort effect being estimated to persist forward, thus highlighting the need for the structural strategy.[4]

These estimates of ability and effort and the persistence of each are vital ingredients in the approach we devise for placing alternative education reforms, including those that alter accountability incentives, on a common footing. Effects on student achievement in the short and longer term associated with ability and effort inform the benefit side of the relevant

[3]To see that one regressor (contemporaneous effort) is inherently a function of the main parameter of interest, note that proficiency-count systems cause educators to make effort decisions by considering how close to the proficiency target a student is expected to score in the absence of any additional effort. But students' predicted scores depend on the degree to which the effort they received in the prior period persists forward, thus rendering contemporaneous effort decisions a function of the persistence rate.

[4]The faster decay we find for effort relative to ability is in line with teachers 'teaching to the test' to some degree – a phenomenon that is often discussed but rarely identified empirically. The fact that a portion of prior effort carries over in scores indicates, however, that effort does have longer-term benefits.

calculation. On the cost side, given that NCLB incentives are sanctions-based, we use our structural model to monetize the value of NCLB sanctions. These costs of incentive reforms can then be placed alongside systematic cost calculations made in the prior literature.

Being able to quantify and compare the benefits and costs of policies affecting teacher ability and effort for the first time allows us to extend the policy discussion in a useful way. The prior literature (notably Hanushek (2009, 2011) and Chetty *et al.* 2014b) has focused on altering the teacher ability distribution as a policy lever to raise student scores. Specifically, policy proposals have featured the notion of replacing teachers whose value-added falls in the bottom five percent of the measured distribution. Building on our finding that incentives matter when measuring the effects of teachers, changing formal incentives constitutes an alternative way of raising student and school performance. Using our framework in an illustrative way, we examine circumstances in which an incentive-based reform can be more cost-effective than the leading ability-based reform considered in the literature, our estimates indicating that it costs 12 percent less on a per-teacher basis for the same longer-run gain.

The remainder of the paper is organized as follows: The next section describes institutional background to our setting, including the accountability programs we rely on for identification and the rich data used in our analysis. Section III presents motivating descriptive evidence; Section IV sets out a framework for analyzing the production of student achievement that provides the basis for our empirical approach; Section V outlines the empirical strategy we employ to decompose ability and effort contemporaneously and also describes the results from that exercise; and Section VI presents our strategy for estimating the persistent effects of effort (versus ability) along with the associated results. Section VII uses the model and estimates to set out a new framework for comparing alternative education policies, which we then use to conduct informative policy comparisons, and Section VIII concludes.

II. Institutional Background and Data

We conduct our analysis in North Carolina, a state that provides significant variation in performance incentives across teachers and schools as well as rich longitudinal data covering all public schools, their teachers and students. We discuss each in turn.

II.A. Accountability Incentives

On the institutional side, the state offers useful incentive variation arising from two separate accountability regimes. The first of these, North Carolina's ABCs of Public Education, was implemented in the 1996-97 school year for all schools serving kindergarten through grade eight. Under the ABCs, each grade from three to eight in every school is assigned a grade-specific growth target, which depends on both average previous student performance and a constant level of expected growth. Based on average school-level gains across all grades in student standardized mathematics and reading scores, the ABCs pay a monetary bonus to all teachers and the principal if a school achieves its overall growth target.[5]

Provisions under NCLB – the second of the accountability regimes operating – were implemented in North Carolina in the 2002-03 school year,[6] following the passage of the federal No Child Left Behind Act in 2001. In contrast to the ABCs' rewards-based approach, NCLB sets penalties for under-performing schools. The program categorizes students into nine subgroups and requires schools to ensure that the percentage of students in each subgroup who achieve proficiency status on the relevant state test meets the state-mandated target. If a school fails to meet any of its subgroup-specific targets, it faces an array of penalties that become more severe over time in the event of repeated failure.[7]

II.B. Data and Descriptive Statistics

Alongside these accountability regimes, North Carolina provides rich longitudinal education data from the entire state, available through the North Carolina Education Research Data Center (NCERDC). These data contain yearly standardized test scores for each student in grades three through eight, encrypted identifiers for students and the teachers who proctor their tests, as well as unencrypted school identifiers. Thus, students can be tracked longitudinally, and linked to a teacher and school in any given year.

Our sample runs from 1997-2005 and is substantial, covering over 2.5 million student-year observations. In terms of performance measures, it includes end-of-grade (EOG) test score performance data for mathematics and reading for all third to eighth grade public

[5] For more detailed descriptions of the ABCs program, see Vigdor (2009) and Macartney (2016).
[6] We will refer to an academic year by its second calendar year. On that basis, NCLB took effect in 2003.
[7] A more detailed description of NCLB can be found in Ahn and Vigdor (2014).

school students in the state, although we focus on students in third to fifth grade, for whom we can construct teacher VA estimates accurately.[8] We also observe a 'pre-test' in grade three, which is written at the beginning of the year and is treated as the grade two baseline test for third graders.

Table 1 provides summary statistics for the unrestricted sample of students. In the analysis below, we construct our ability and effort measures for each teacher by using individual student test scores. These are measured on a developmental scale, designed so that each additional point represents the same amount of knowledge gained, irrespective of the baseline score and school grade. Both the mathematics and reading scores in the table show a monotonic increase across grades, consistent with knowledge being accumulated in those subjects over time. The test score *levels* are relevant under NCLB, which requires that a given proportion of each of the nine student subgroups (referred to above) exceeds a target score on standardized tests.

The longitudinal nature of the data set enables us to construct growth score measures for both mathematics and reading, based on within-student gains. Those gains are positive, on average, in both subjects across grades, though the largest gains occur for both subjects 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 includes information about individual students' gender, race, disability status, limited English-proficiency classification, free lunch eligibility, and grade progression. In the aggregate, about 40 percent of students are minorities (non-white), 6 percent are learning-disabled, only 3 percent are limited English-proficient, and 44 percent are eligible for free or reduced-price lunch. Around 25 percent of students have college-educated parents, and very small fractions of students repeat a grade. These demographic characteristics serve as control variables in our analysis below.

Given our interest in exploring the separate effects of teacher ability and effort, we need to match students in the EOG files to their teachers in an accurate way in any given year. We construct the sample used in our analysis by following previous studies that use the NCERDC data, restricting attention to students in third through fifth grade (as mentioned),

[8] See the discussion at the end of the section for an explanation.

Table 1: Student-Level Summary Statistics

	Mean	Std. Dev.	Obs.
Performance Measures			
Math Score			
Grade 3	144.67	10.67	905,912
Grade 4	153.66	9.78	891,971
Grade 5	159.84	9.38	888,469
'Future' Math Score			
Grade 6	167.16	11.01	739,386
Grade 7	172.61	10.70	617,669
Grade 8	175.79	11.36	503,091
Math Growth			
Grade 3	13.88	6.30	827,738
Grade 4	9.20	6.02	817,240
Grade 5	6.92	5.35	815,602
Reading Score			
Grade 3	147.03	9.33	901,235
Grade 4	150.65	9.18	887,153
Grade 5	155.79	8.11	883,689
'Future' Reading Score			
Grade 6	157.07	8.66	737,192
Grade 7	160.76	8.00	616,384
Grade 8	163.32	7.56	502,229
Reading Growth			
Grade 3	8.20	6.71	838,387
Grade 4	3.85	5.58	811,890
Grade 5	5.49	5.22	810,216
Demographics			
College-Educated Parents	0.25	0.43	2,757,648
Male	0.51	0.50	2,778,454
Minority	0.40	0.49	2,7767,29
Disabled	0.06	0.24	2,778,635
Limited English-Proficient	0.03	0.17	2,778,623
Repeating Grade	0.02	0.13	2,778,734
Free or Reduced-Price Lunch	0.44	0.50	1,998,653

Notes: Summary statistics are calculated for all third through fifth grade student-year observations from 1997 to 2005. The free or reduced-price lunch eligibility variable is not available prior to 1999. Math scores are measured on different scales before and after 2001. We are able to convert second edition scale scores to their first edition counterparts for all tests except the grade three pre-test (the 'grade two' test). Thus, all level and gain math score summary statistics are expressed on the first edition scale except grade three gains, which are calculated using first edition scores prior to 2001 and second edition scores for 2001 onwards. 'Future' math and reading scores are the scores we observe for our sample of third to fifth grade students when they are in sixth, seventh, and eight grades. We use these scores when measuring the persistent effects of teacher ability and effort. We do not follow students past 2005, as the math scale changes again in 2006 but no table to convert scores back to the old scale was created by the state.

where the teacher recorded as the test proctor tends to be the teacher who taught the students throughout the year. We also follow Clotfelter, Ladd, and Vigdor (2006) and subsequent research by only counting a student-teacher match as valid if the test proctor in the EOG files teaches a self-contained class for the relevant grade in the relevant year and if at least half of the tests administered by that teacher are for students in the correct grade. Special education and honors classes are excluded from the analysis, but we retain students who repeat or skip grades.

III. Motivating Descriptive Evidence

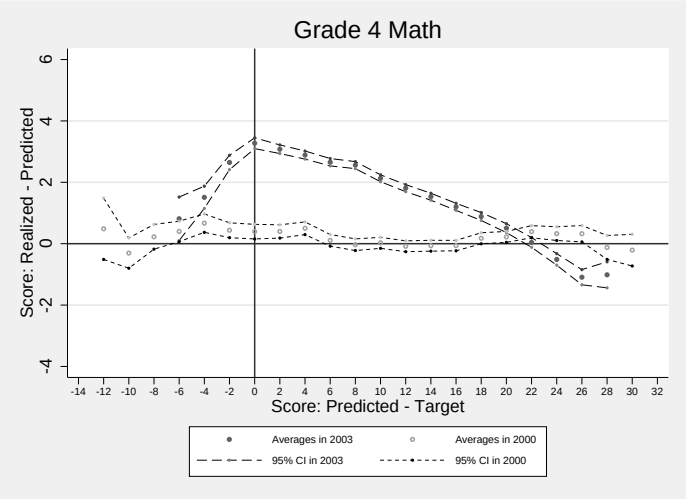
In this section, we present suggestive evidence to motivate the subsequent empirical analysis. First, we provide clear evidence of an incentive response consistent with there being the targeted effort increase, as expected under proficiency count incentive schemes; second, we provide correlational evidence suggesting that teacher value-added is responsive to incentives.

Our strategy for uncovering teacher effort draws heavily on the introduction of NCLB in 2003, treating that as an exogenous shock to educators' performance incentives. It has already been well-established that proficiency-count systems like NCLB provide educators with strong incentives to direct resources to students who are on the margin of passing relative to the fixed target, potentially at the expense of those in the tails of the predicted test score distribution.[9] Thus, the introduction of NCLB should give rise to an inverted U-shaped increase in test scores, peaking where students are directly at the margin of passing.

That prediction is captured empirically in Figure 1, which shows that this is exactly the response that occurred in North Carolina when NCLB was introduced in 2003. The figure indicates that gains over predicted test scores in 2003 were highest for students predicted to score close to the test-score proficiency threshold. Gains over predicted scores are decreasing as one moves further away from that threshold, consistent with the standard predicted effort responses to proficiency-count programs.

In contrast, also shown in the figure, students experienced no gain over their predicted scores at any point of the predicted score distribution in the *pre*-NCLB period. The flat profile pre-NCLB lends credence to the notion that the incentive shock was exogenous: distance from the passing threshold has essentially zero predictive power.

[9] See, for example, Reback (2008), Ladd and Lauen (2010), and Neal and Schanzenbach (2010).



Notes: This is Figure 3 from Macartney *et al.* (2015). The figure is constructed as follows: In each year, we calculate a predicted score for each grade four student and then subtract off the known proficiency score target from this prediction – the horizontal axis measures the difference. We then group students into 2-point width bins on the horizontal axis. Within each bin, we calculate the average (across all students) of the difference between students' realized and predicted scores. The circles represent these bin-specific averages: the solid circles represent year-2003 averages; the hollow circles are year-2000 averages. The figure also shows the associated 95 percent confidence intervals for each year. Standard errors are clustered at the school level.

Figure 1: Inverted-U Response to NCLB

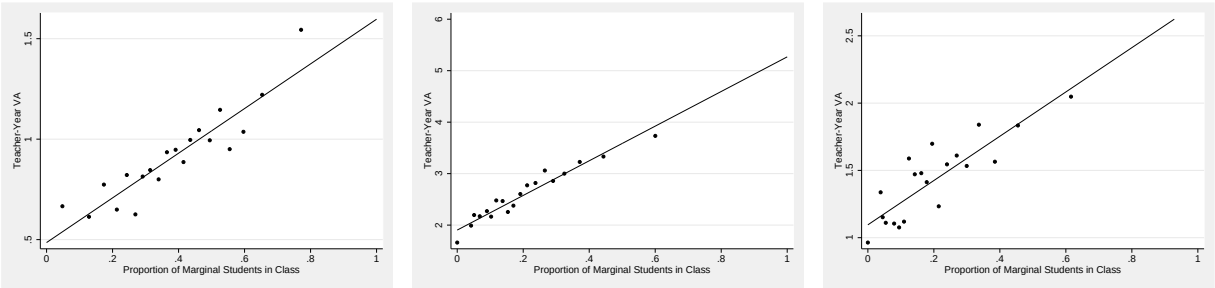
Next, building on the notion that NCLB provided differential incentives to exert greater effort depending on how marginal students were relative to the fixed score target, we provide evidence indicating that teacher-year fixed effects, which are commonly used to measure teacher effectiveness, actually covary with a simple proxy for NCLB incentive strength in 2003.[10] We start with the proxy for incentive strength, defining a student as 'marginal' if she is predicted to score within $+/-4$ developmental scale points of the proficiency cutoff,[11] and calculate the fraction of students in each classroom who are marginal in that sense. Then, since teacher-year VA represents *average* residual student test score gains within a classroom, teacher-year VA should be an increasing function of *average* student NCLB incentive strength within the relevant classroom.

This is also what we find, as Figure 2 shows. Teacher-year fixed effects depend positively on the proportion of marginal students within a classroom. The relationship is significant (at the one percent level) and positive in each grade in 2003, with a one standard deviation increase in the proportion of marginal students within a classroom being associated with 7, 17, and 11 percent standard deviation increases in teacher-year VA in third, fourth, and fifth

[10]The estimation of teacher-year fixed effects is discussed in much greater detail in Section V below.

[11]The results are robust to various alternative choices of cutoff for defining marginal students.

grade, respectively. These raw-data patterns suggest NCLB caused teachers to exert more effort in a manner guided by the prevailing incentives. In contrast, in the pre-NCLB period, we would expect there to be no relationship between the proportion of marginal students in a classroom and teacher VA, as we document in Section V.[12]



(a) Third Grade (b) Fourth Grade (c) Fifth Grade

Notes: This figure depicts the relationship between teacher-year VA measures and the fraction of marginal students within a classroom in the year 2003. To construct the figure, we first group teacher-year observations into 20 equally-sized (vingtile) bins of the distribution of the fraction of marginal students on the horizontal axis. Within each bin, we calculate the average proportion of marginal students and the average teacher-year VA estimate. The dots in each panel represent these averages in 2003. The lines represent the associated linear fits, estimated using the underlying teacher-year data.

Figure 2: Teacher-Year Fixed Effects versus the Proportion of Marginal Students

IV. Theory

In this section, we present a model that serves to motivate our approach to identifying teacher ability and effort separately as well as estimating the persistence of each – issues we address later in the section.

IV.A. The Environment

The model has two main elements: an education production technology and the teacher effort decision, made in light of prevailing accountability incentives. Teacher and student assignments are taken as given, and we suppress several other aspects of education production to ease exposition, including student effort, student peer effects, and parental help. In Section IV.C, we discuss the implications of allowing for a more general setup – one that allows for alternative specifications of the technology as well as other inputs.

[12]There, we argue that more than simple correlational plots are required to account for a confounding negative correlation between marginal student presence and teacher ability. In that section, we estimate incentive-invariant teacher ability and account for potentially differential sorting of students to teachers in 2003, along with the effects of teacher experience and random classroom shocks to performance to ensure that the relationship in Figure 2 indeed reflects a teacher effort response.

IV.A.1 Production Technology

We begin with the education production function. As the exact specification of the technology is unknown and many inputs are unobserved, researchers typically limit analyses to inputs that are observed in school administrative data sets. Furthermore, most studies examine the effects of contemporaneous inputs, accounting for the history of past inputs by controlling for prior test scores. In contrast, our analysis focuses on both the contemporaneous and dynamic effects of two unobserved inputs, namely teacher ability and effort.

To fix ideas, we abstract from other inputs and follow convention in the literature by making

Assumption 1: The education production technology is linear.

Specifically, we consider the following representation for the technology, which makes explicit how each of the two inputs affects student learning, both contemporaneously and over time:

$$y_{ijgst} = \sum_{0 \leq t' \leq t} \left[\gamma^a_{t,t'} a_{j(i,t')} + \gamma^e_{t,t'} e_{j(i,t')} \right] + \nu_{igjst} \,. \tag{1}$$

Equation (1) describes the test score of student student i, assigned to teacher j in grade g at school s and time t. Here, $a_{j(i,t')}$ is the ability of the teacher whom student i was assigned to at time t', $e_{j(i,t')}$ is the effort choice of that teacher at that time, and ν_{igjst} is an error term.[13]

Our empirical goals are to separately identify teacher ability and effort, $a_{j(i,t')}$ and $e_{j(i,t')}$, and their effects on test scores, $\{\gamma^a_{t,t'}, \gamma^e_{t,t'}\}_{t'<t}$. To simplify the analysis, we equate the contemporaneous effects of teacher ability and effort, assuming $\gamma^a_{t,t} = \gamma^e_{t,t} = 1$. Separate identification of ability and effort will require observing a teacher in at least two different time periods in which effort incentives differ, a condition satisfied in our setting by the introduction of NCLB into a state with pre-existing incentives.

To clarify the link between the technology just outlined and our subsequent empirical analysis, one can first multiply the prior score by γ, which represents the rate at which the stock of knowledge accumulated up to period $t-1$ persists to affect current test scores;[14] this parameter is a composite measure of the persistent effects of teacher ability, teacher effort,

[13] As we discuss in Section IV.C, neither the omission of other education inputs nor the assumed linearity of effort and ability are critical to our empirical analysis.

[14] Todd and Wolpin (2003) provide a detailed discussion of specifications for the production technology that render the contemporaneous test score a function of prior test scores.

and random shocks to performance. Next, subtract the result form both sides of the test score equation:

$$\begin{aligned} y_{ijgst} - \gamma y_{i,j'g-1s't-1} = a_{j(i,t')} + e_{j(i,t')} \\ + \sum_{0 \leq t' \leq t-1} \left[(\gamma^a_{t,t'} - \gamma\gamma^a_{t-1,t'}) a_{j(i,t')} + (\gamma^e_{t,t'} - \gamma\gamma^e_{t-1,t'}) e_{j(i,t')} \right] \\ + \left(\nu_{igjst} - \gamma \nu_{ijg-1js't-1} \right). \end{aligned} \tag{2}$$

The resulting expression for test scores, with the relevant (simplifying) relabeling, is

$$y_{ijgst} = \gamma y_{ij'g-1s't-1} + a_j + e_{ijgst} + \epsilon_{ijgst}. \tag{3}$$

The simplified production technology thus implies that the test score of student i depends on his or her prior score (or accumulated stock of knowledge), his or her teacher's ability, the effort the teacher chooses to devote to the student, and a random shock to performance. We will use this as the basis for estimation below.

IV.A.2 Effort Decision

In this subsection, we first describe teachers' effort decisions under the ABCs, showing that the program results in relatively uniform effort incentives across teachers within a school. We then show that NCLB's introduction generates systematic variation in incentives across teachers, providing a means to identify teacher effort separately from teacher ability.

ABCs: The ABCs program sets growth targets that are grade- and subject-specific for each school and then aggregates across all grade-subject pairs within a school to form a school-level growth score. Let G_s denote the highest grade served by school s. In the ABCs legislation, average growth targets in a subject-grade are set as a linear function of students' prior scores. Let α denote the (pre-determined) coefficient that multiplies student prior scores to form these targets.[15] School s passes the ABCs when the sum of the differences

[15] Using a single multiplicative coefficient, α, and one prior score in the ABCs target is a simplification. In practice, the ABCs program sets targets using both student prior math and reading scores, and individual multiplicative coefficients for each. We note, however, that the actual targets are *linear* in the prior scores, which is all that is needed for our theoretical results.

between average and target scores in each grade is greater than zero:

$$\sum_{g=3}^{G_s} \sum_{\{i:\ i \in g_{st}\}} \frac{y_{isgt} - \alpha y_{ij'g-1s't-1}}{N_{gst}} \geq 0. \tag{4}$$

Here, the sum is taken over all students who are in grade g in school s at time t, and N_{gst} is the number of students in that grade in school s at time t.

Substituting the production technology from equation (3) into equation (4) yields

$$\sum_{g=3}^{G_s} \bar{\epsilon}_{gst} \geq \sum_{g=3}^{G_s} \left((\alpha - \gamma)\bar{y}_{g-1st-1} - \bar{a}_{gst} - \ \bar{e}_{gst} \right), \tag{5}$$

where the upper bars denote school-grade-year-specific averages. We let $F(\cdot)$ denote the cumulative density function of the school-average error term, $\sum_{g=3}^{G_s} \bar{\epsilon}_{sgt}$. To simplify the exposition, we make

Assumption 2: The cost of teacher effort does not depend on student characteristics and is given by the strictly increasing and convex function, $c(e_{ijgst})$.

Specifically, we assume that an additional unit of effort is associated with the same additional cost, regardless of whether it is directed toward a high-performing student or a low-performing student. In Section IV.C, we discuss the implications of relaxing this assumption, showing that it is not critical to our main analysis.

Letting N_{st} denote the total number of students attending the school, we model the school as a centralized decision-maker, choosing a sequence of effort levels across all students $\{e_{ijgst}\}_{i=1}^{N_{st}}$ in each year to maximize the following expected-utility payoff:

$$b^{abcs}\left(1 - F\left(\sum_{g=3}^{G_s}\left((\alpha - \gamma)\bar{y}_{g-1st-1} - \bar{a}_{gst} - \ \bar{e}_{gst}\right)\right)\right) - \sum_{g=3}^{G_s} \sum_{\{i:\ i \in g_{st}\}}^{N_{st}} c(e_{ijgst})\,. \tag{6}$$

Here, b^{abcs} represents the financial bonus that is paid to all teachers and the school principal when the school passes the ABCs. To further simplify the notation, we define $\Omega_{st} \equiv \sum_{g=3}^{G_s}\left((\alpha - \gamma)\bar{y}_{g-1st-1} - \bar{a}_{gst} - \ \bar{e}_{gst}\right)$. The first-order condition for the optimal effort devoted to any student i is given by

$$b^{abcs} f(\Omega_{st}) \frac{1}{N_{gst}} = c'(e_{ijgst}) \,, \tag{7}$$

where $f(\cdot)$ is the probability density function of $\sum_{g=3}^{G_s} \bar{\epsilon}_{sgt}$. Lemma 1 establishes a useful property characterizing ABCs-related effort.

Lemma 1 *Assume constant grade sizes within each school, such that $N_{g'st} = N_{gst}$ for all $g', g \in G_{st}$. Then each student receives the same amount of effort within a school: ABCs effort varies only at the school level.*

The proof of Lemma 1 follows immediately from the first-order condition in equation (7), which shows that both the marginal benefit of effort and the marginal cost are the same for all students within a given school.[16] Since in practice, student characteristics do not vary much over time within schools, we follow Macartney (2016) and assume that schools reach steady-state levels of ABCs effort, implying that all effort variation occurs across schools and is driven by variation in the steady-state likelihood of ABCs success, denoted by $f(\Omega_s^*)$.

NCLB: Starting in 2003, schools face incentives under NCLB, in addition to the ABCs. NCLB sets test score proficiency targets that are fixed across students in a given grade; and it requires that a predetermined percentage of students within a school achieve proficiency status. We abstract from NCLB's subgroups here, instead assuming that only the school-level pass rate is relevant for success or failure under NCLB.[17]

Letting $y_g^{T,nclb}$ denote the test score proficiency target in grade g under NCLB, the probability that student i reaches proficiency status under NCLB may be written as

$$1 - H(y_g^{T,nclb} - \gamma y_{ig-1s't-1} - a_j - e_{igst}), \tag{8}$$

[16] The result depends on there being no interaction between student characteristics and teacher effort in either the production technology or the cost function. While allowing for such interactions may overturn the result, it does not invalidate the empirical analyses that follow, as we discuss in Section IV.C.

[17] We do not model subgroups directly, which is a reasonable first-order approximation to the incentives we exploit: holding fixed whether or not a given subgroup is held accountable under NLCB, the marginal benefit of effort for educators within that subgroup is always higher for students who are predicted to score near the proficiency threshold, irrespective of whether the subgroup is actually accountable. While students within accountable subgroups receive more effort, schools face incentives to devote effort to all marginal students, as marginal students who are not members of accountable subgroups in the current year may be members of accountable subgroups in future years.

after using the production technology in equation (3), and letting $H(\cdot)$ represent the cumulative density function of ϵ_{igst}. Aggregating across all students, the school's expected pass rate becomes

$$R_{st} = \sum_{g=3}^{G_s} \sum_{\{i:\ i \in g_{st}\}} \frac{1 - H(y_g^{T,nclb} - \gamma y_{ij'g-1s't-1} - a_j - e_{igst})}{N_{st}}. \tag{9}$$

We follow Neal and Schanzenbach (2010), assuming that the school's NCLB payoff is increasing in the fraction students who are proficient (independent of how that fraction compares with the required school-level proficiency rate) and approximating the total number of students who pass as the expected number who pass. Since NCLB does not pay out a financial bonus when schools succeed, nor does it apply a financial sanction when schools fail, we cannot write down an explicit benefit of having a higher school-level pass rate under NCLB. Instead, we assume that educators attach some value to higher performance, letting $\Psi(\cdot)$ denote the strictly increasing and concave reward function that takes the fraction of students who pass as its argument.[18]

Combining the model of NCLB with that of the ABCs above, the school's objective function in 2003 becomes

$$U_{st} = b^{abcs}\big(1 - F(\Omega_s)\big) + \Psi\big(R_{st}\big) - \sum_{g=3}^{G_s} \sum_{\{i:\ i \in g_{st}\}} c(e_{ijgst}), \tag{10}$$

and the first-order condition for the effort devoted to any student i is now

$$b^{abcs} f(\Omega_s) \frac{1}{N_{gst}} + \Psi'\big(R_{st}\big) \frac{h(y_g^{T,nclb} - \gamma y_{ij'g-1s't-1} - a_j - e_{igst})}{N_{st}} = c'(e_{ijgst}). \tag{11}$$

This equation makes clear how the effort devoted to each student within a school now has both an individual and common component. The individual component comes from NCLB given that. Under this scheme, a student's position in the student ability distribution determines how much effort he or she receives under this scheme, as dictated by the function $h(y_g^{T,nclb} - \gamma y_{ij'g-1s't-1} - a_j - e_{ijgst})$.

To understand which students receive the most additional effort under NCLB, it is helpful

[18] Neal and Schanzenbach (2010) model the expected *fail* rate and thus have a strictly increasing and convex *penalty* function.

to decompose teacher effort implied by equation (11), writing the total effort received by student i as the sum of ABCs- and NCLB-related effort, $e_{ijgst} = e_s^{*abcs} + e_{ijgst}^{nclb}$. Here, e_s^{*abcs} represents steady-state ABCs effort, which does not vary within schools, and e_{ijgst}^{nclb} represents student-specific NCLB effort. We denote the score that educators predict students to achieve without any additional NCLB response as $\hat{y}_{ijgst} \equiv \gamma y_{ij'g-1s't-1} + a_j + e_s^{*abcs}$. The following proposition establishes that students with intermediate predicted scores are those who receive the most additional effort, while students at the high and low extremes of the predicted score distribution receive comparatively less effort.

Proposition 1 *Define* $\pi_i(\hat{y}_{ijgst}) = y_g^{T,nclb} - \hat{y}_{ijgst}$ *as the distance between the NCLB test score proficiency target and student i's predicted score and assume that* $h(\cdot)$ *has a unimodal distribution. Then there exists an intermediate value* $\hat{y}^*$*, such that* $\pi_i(\hat{y}^*)$ *is relatively small in absolute terms, which results in maximal NCLB effort within a given school. Further, effort is decreasing as* $\hat{y}$ *decreases below or increases above* $\hat{y}^*$*.*

Proof See Appendix. ■

Proposition 1 implies that the students for whom the distance between the proficiency target and predicted score is relatively small (in absolute terms) receive the most additional effort under NCLB. Students who are predicted to score far above the threshold are likely to pass without any additional help, while students who are predicted to score far below require a prohibitively costly amount of effort to change their expected proficiency status. Thus, the marginal net benefit of extra effort for educators is highest among students predicted to be on the margin of passing. NCLB therefore generates variation in effort that occurs across students within schools, in contrast to the ABCs, where effort varies mainly across schools.

While the result illustrated in Proposition 1 is not new to the literature,[19] it serves in our empirical analysis to identify teacher ability and effort separately and to estimate the persistent effects of each. Next, we use the model to provide intuition for our identification strategies.

[19]For example, similar results are established in Reback (2008) and Neal and Schanzenbach (2010).

IV.B. Using the Model to Motivate Identification Strategies

IV.B.1 Intuition for Separating Ability and Effort

Our first goal is to separately identify teacher ability and effort. Here, the model suggests a natural way to decompose traditional teacher-year VA measures, representing average residual test scores across all students within a given classroom, into teacher ability and effort components. To explain the intuition, we make an additional assumption about the cost of effort:

Assumption 3: The cost of effort is given by $c(e) = \frac{d}{2}e^2$, where $d > 0$.

In this case, the VA estimate for teacher j in 2003, q_{j2003}, may be written as

$$\begin{aligned} q_{j2003} &= \sum_{i=1}^{N_{j2003}} \frac{y_{ijgst} - \gamma y_{ij'g-1s't-1}}{N_{j2003}} \\ &= \sum_{i=1}^{N_{j2003}} \frac{a_j + e_{ijgst} + \epsilon_{ijgst}}{N_{j2003}} \\ &= a_j + \frac{b^{abcs} f(\Omega_s)}{dN_{gst}} + \frac{\Psi'\big(R_{st}\big)}{dN_{st}} \sum_{i=1}^{N_{j2003}} \frac{h(\pi_i - e_{igst})}{N_{j2003}} + \sum_{i=1}^{N_{j2003}} \frac{\epsilon_{ijgst}}{N_{j2003}}, \end{aligned} \tag{12}$$

where the third equality follows from substituting the first-order condition for effort (equation (11)) into the previous line. A teacher's VA measure in 2003 thus captures the teacher's incentive-invariant ability level, a_j, baseline ABCs effort, average NCLB-related effort across all students in the teacher's class, and classroom average noise in test scores.

Proposition 1 showed that NCLB effort will be highest among students for whom the distance between the proficiency target and predicted score, π_i, is relatively small in absolute terms. At the same time, students who are predicted to score relatively far away from the target (in either direction) receive comparatively less effort. Equation (12) shows that these student-specific effort incentives influence teacher VA measures directly, as average classroom effort is a function of average student incentives: $\sum_{i=1}^{N_{j2003}} \frac{h(\pi_i - e_{igst})}{N_{j2003}}$.

Since students who are predicted to be on the margin of passing (scoring relatively close to the proficiency target) receive the most additional effort under NCLB, it follows that average classroom effort will be increasing in the classroom fraction of these marginal

students. In Section V, we use this fact to generate exogenous variation in effort incentives across teachers, allowing us to identify teacher ability and effort separately by comparing teacher VA performance in 2003 to that in prior years when NCLB incentives do not operate.

IV.B.2 Intuition for Estimating the Persistence of Effort

Our second research question involves estimating how the effects of ability and effort persist. Focusing on effort, in 2004 we allow prior-period NCLB effort to persist at a potentially different rate than pre-NCLB test score inputs, which persist at the common rate γ, noting that we have no way of separating out different components prior to NCLB's introduction. Test scores in 2004 are written as

$$y_{ijgst} = \gamma(y_{ij'g-1s't-1} - e^{nclb}_{ij'g-1s't-1}) + a_j + \tilde{e}^{abcs}_s + e^{nclb}_{ijgst} + \gamma^e e^{nclb}_{ij'g-1s't-1} + \epsilon_{ijgst}, \qquad (13)$$

where γ^e denotes the persistence rate of NCLB effort from 2003. ABCs effort is denoted as $\tilde{e}^{abcs}_s$ to indicate that it is no longer necessarily the steady-state level of effort chosen by school s prior to NCLB, owing to NCLB's potential disruption of ABCs targets.

There are two key challenges to identifying the persistence of effort, γ^e, empirically. First, NCLB-related effort decisions in 2004 are a function of NCLB effort from 2003 and the persistence rate. To see this, note that the student-specific term in the first-order condition for effort becomes $h\big(y^{T,nclb}_g - \gamma(y_{ij'g-1s't-1} - e^{nclb}_{ij'g-1s't-1}) - a_j - \tilde{e}^{abcs}_s - e^{nclb}_{ijgst} - \gamma^e e^{nclb}_{ij'g-1s't-1}\big)$ in 2004. The persistence rate of prior-year effort determines student human capital in 2004 and, correspondingly, how close to the test-score proficiency target a student is expected to score. Educators thus make contemporaneous student effort decisions in 2004 based in part on the degree to which they believe prior effort will persist.

Second, the persistence rate of prior effort may also potentially affect schools' ABCs incentives. In particular, writing the probability of a school passing the ABCs in 2004 as

$$1 - F\left(\sum_{g=3}^{G_s}\left((\alpha-\gamma)\bar{y}_{g-1st-1} + (\gamma-\gamma^e)\bar{e}^{nclb}_{g-1st-1} - \bar{a}_{gst} - \bar{\tilde{e}}^{abcs}_{gst} - \bar{e}^{nclb}_{gst}\right)\right), \qquad (14)$$

then if $\gamma > \gamma^e$, NCLB effort from the prior year decreases the probability of ABCs target attainment in 2004, holding all else equal. This is due to the discrepancy between the actual

persistence of effort and the rate assumed in setting the ABCs target. The target takes the full prior score into account, not discriminating between potentially differential persistence rates of the inputs that contribute to that score. If effort persists at a *lower* rate than non-effort inputs, however, test scores in 2004 increase at a lower rate than they would in the pre-NCLB period, when non-effort inputs are responsible for all test score gains. Since the ABCs target holds schools to the same standard as in the pre-reform period, the same *level* of test score improvement from the year $t-1$ makes it more difficult for schools to reach contemporaneous ABCs targets in 2004 relative to pre-NCLB years.

It is also clear that *average* 2003 NCLB effort is the key determinant of the distortion to 2004 school-level ABCs incentives. There is no distortion only when effort persists at the same rate as all non-effort inputs, such that $\gamma = \gamma^e$. Otherwise, responses to NCLB in 2003 have direct implications for schools' ABCs effort choices in 2004, as schools respond to the changed likelihood of meeting ABCs targets. It is important to note, however, that the model implies that ABCs effort continues to be the same across all students within a given school. In Section VI, we make use of these observations as a way to control for the effects of ABCs effort while estimating the persistence of NCLB effort from 2003.

IV.C. The Implications of Considering a More General Model

Non-Linearities: Interactions between Student Characteristics and Teacher Effort

In the model, teacher effort does not interact with student ability in the production function and that the costs of effort do not depend on student characteristics. Neither assumption need hold true in practice for our empirical analyses to be valid. The key claim from our model (Proposition 1) is that NCLB creates the strongest incentives to exert additional effort to marginal students, *relative to the effort identical students would have received prior to NCLB*. While interactions in the production technology – for example, complementarities between teacher effort and student ability – or heterogenous effort costs across students change relative effort levels *across* students, such specifications still imply that marginal students receive the greatest amount of *additional* effort. One can alternatively conceptualize the idea as marginal students receiving the largest within-student increase in effort, relative to the counterfactual scenario without NCLB. This result does not depend on the linearity assumed above, as it is also predicted by data-generating processes with more complicated

interaction terms.

Models with interactions between student characteristics and teacher effort do affect the result in Lemma 1, which shows that ABCs-related effort is uniform across students within a school. Uniform effort levels imply that there are no effort differences across teachers within schools prior to NCLB, allowing us to measure the NCLB-related increase in effort relative to a common baseline across all teachers. If the data-generating process involves interactions between student characteristics and teacher effort, then the sorting process of students to teachers implies that some teachers start from different baseline effort levels than others. Since we estimate NCLB effort below by controlling for a composite measure of teacher ability and baseline effort, we require in this case that students are not *differentially* sorted to teachers after NCLB is introduced.[20] Without differential sorting, ABCs incentives remain constant within-teacher, allowing us to identify NCLB effort responses even under more complicated data-generating processes.

Other Inputs in Education Production

In focusing on teacher ability and effort, the model ignores many other inputs in education production, including student effort, student peer effects, and parental help. As we explain below, NCLB effort is estimated in a difference-in-differences type of framework, implying that the effects of such inputs are eliminated if the inputs do not adjust in response to NCLB incentives. It may be the case, however, that some inputs do adjust in response to teachers changing their effort decisions under NCLB. For example, parents may decrease the amount of help they give their children when teachers increase effort, or students may try harder when they notice teachers giving them more attention. In these cases, our estimates reflect such adjustments as the general effects of teacher effort. While we cannot isolate the partial effect of effort in such instances, we do still estimate the policy-relevant parameters, as these inputs will adjust in response to effort changes under any proposed accountability scheme similar to NCLB.

[20] We test this assumption empirically below.

V. Separating Contemporaneous Effort and Ability

In this section, we describe our identification strategy for decomposing a teacher's contribution to her students' test scores into ability and effort components, then present our findings.

V.A. Identification Strategy

As a starting point, it is useful to place the decomposition we have in mind alongside the output generated by standard teacher VA methods. Those methods construct average residual test score gains across all classrooms taught by a given teacher, relying on several observations for the same teacher over time to minimize the influence of noise. VA estimates for a given teacher thus reflect that teacher's incentive-invariant ability and the average effort she devotes to all of her students over many years.

If, on the one hand, effort incentives exhibit substantial within-teacher variation, VA methods cleanly identify teachers' incentive-invariant ability levels but the variation in effort decisions over time implies that the effects of effort are averaged out. When, on the other, incentives do not vary much over time for a given teacher, VA estimates reflect a composite of both teacher ability and average effort. In neither case do VA measures distinguish explicitly between the two components of teacher effectiveness.

Our approach to doing so involves a three-step procedure, with an emphasis on placing very limited structure on the decomposition problem. By way of overview, we first compute teacher-year fixed effects for each teacher from 1997 to 2005, a step largely in keeping with existing approaches taken in the literature. Second, we use pre-reform data to identify the sum of incentive-invariant ability and pre-existing baseline (ABCs) effort using the Empirical Bayes estimator of teacher VA (Kane and Staiger, 2008; Chetty *et al.*, 2011). Third, we estimate NCLB-induced teacher effort, using the estimated teacher fixed effects from 2003 along with estimates of the sum of teacher ability and baseline effort, and the fraction of students in a teacher's classroom deemed 'marginal' with respect to the NCLB target.[21] We now describe each step in greater detail.

1. Teacher-Year Fixed Effects: When estimating teacher VA, we regress contemporaneous math scores on cubics in prior math and reading scores and several other student

[21] Here, binary and continuous definitions can be applied. As we show, the underlying results are insensitive to the specific definition chosen.

characteristics, in addition to the teacher-year fixed effects that are our primary focus. In doing so, we follow recent studies.

While test scores that are standardized (usually at the grade-year level) are the main outcome measure in the literature, we opt to measure test scores on a developmental scale instead. On the one hand, standardizing test scores guards against changes in testing regimes over time; on the other, de-meaning effectively removes the effects of aggregate changes in performance incentives. As our primary interest is assessing how teacher effort affects student learning, we prefer to preserve all incentive-related performance variation over time and therefore measure test scores on a developmental scale throughout our analysis. Here, we rely on the careful psychometric design of these scales, which ensures that one can track improvements or declines in learning as students progress through school.

We define a set of controls, x_{ijgst}, to include student race, gender, disability status, limited English-proficiency classification, parental education, and an indicator for grade repetition. The inclusion of these variables helps mitigate any bias stemming from non-random student-teacher matching, as they are likely correlated with innate student ability and previous teacher assignments. A teacher-year fixed effect is calculated for a teacher only if she has greater than seven but fewer than forty students in her class with valid test scores and demographic variables in the relevant year.[22]

To estimate teacher-year fixed effects, we use the full sample from 1997 to 2005 and run the following grade-specific regressions (for third, fourth and fifth grades):[23]

$$y_{ijgst} = f(y_{ij'g-1s't-1}) + q_{jt} + x'_{ijgst}\beta + \epsilon_{ijgst}\,, \tag{15}$$

[22]In doing so, we follow the existing literature that has used the North Carolina data. We exclude a student from the value-added analysis if any of the following criteria are satisfied: (1) multiple scores for current or lagged end-of-grade (EOG) math or readings tests; (2) EOG scores corresponding to two or more teachers in a given year; (3) EOG scores corresponding to two or more grades in a given year; (4) or EOG scores corresponding to two or more schools in a given year.

[23]We estimate the teacher-year fixed effects using Stata's 'areg' command, which solves the co-linearity problem arising from including a fixed effect for each teacher-year pair by estimating each teacher-year effect relative to an arbitrarily selected constant (equal to the average teacher-year fixed effect in the case of constant class size). Since the regressions are grade-specific, the teacher-year effects provide a ranking of teachers over time (1998 to 2005) within a given grade. For more details, see McCaffrey *et al.* (2012).

obtaining the teacher-year fixed-effects estimates as

$$\begin{aligned}\hat{q}_{jt} &= \sum_{i=1}^{n(j,t)} \frac{y_{ijgst} - \hat{f}(y_{ij'g-1s',t-1}) - x'_{ijgst}\hat{\beta}}{n(j,t)} \\ &= a_j + \frac{b^{abcs} f(\Gamma_{st})}{dN_{gst}} + 1_{nclb}\frac{\Psi'\left(R_{st}\right)}{dN_{st}} \sum_{i=1}^{N_{j2003}} \frac{h(y_g^{T,nclb} - \gamma y_{ij'g-1s't-1} - a_j - e_{igst})}{N_{j2003}} + \bar{\epsilon}_{jt} \\ &= a_j + \underline{e}_j + 1_{nclb} e_{jt}(m_{jt}) + \bar{\epsilon}_{jt} \qquad (16)\end{aligned}$$

where the second equality follows from equation (12) in the model above and 1_{nclb} is a binary variable that switches on to indicate the post-NCLB period. Teacher-year fixed effects consist of incentive-invariant ability and, in North Carolina, they also consist of baseline ABCs effort, which is denoted as $\underline{e}_j$ throughout the empirical analysis. From Lemma 1, ABCs effort does not vary across teachers within a school, implying that VA methods will capture variation in incentive-invariant ability prior to NCLB.

Once NCLB takes effect in 2003, the composition of students in a teacher's classroom helps to determine effort incentives. In particular, since schools want to devote more effort to students who are predicted to score near the test score proficiency standard, the fraction of these marginal students in the classroom, m_{jt}, is likely to influence teacher performance in 2003 and beyond.

While teacher performance in 2003 is a function of classroom marginal student presence, we cannot estimate the effect of NCLB incentives by simply regressing teacher-year fixed effects on the fraction of marginal students within a classroom because students are sorted to teachers in a systematic way. Specifically, the proficiency cutoff is set at a relatively low level in North Carolina, implying that marginal students are typically low-performing students, who tend to be sorted to low-ability teachers. Therefore, teacher ability and marginal student presence within a classroom are negatively correlated, implying that the estimated effect of NCLB incentives (as measured by marginal student presence across classrooms) is downward biased when one does not control for teacher ability. We address this issue by estimating teacher incentive-invariant ability in the pre-NCLB period and controlling for it directly throughout the analysis.

2. Incentive-Invariant Ability and Baseline Effort: Consistent with much of the existing literature, we assume incentive-invariant ability is fixed over time, conditional on

teacher experience.[24] We assume baseline ABCs-related effort is fixed also, as ABCs incentives only vary at the school level under the plausible conditions outlined above, and most schools reach steady-state levels of effort prior to NCLB.[25] For each teacher j, we estimate the combination of her fixed ability and baseline effort by employing the Empirical Bayes ("EB") estimator of teacher VA (see Kane and Staiger, 2008; Chetty *et al.* 2011) in the pre-NCLB period. Specifically, we estimate the following pooled specification across all grades and years from 1997 to 2002,[26] in which we regress test scores on grade-specific cubic polynomials of prior scores, indicators for student ethnicity, gender, limited-English proficiency, disability status, parental education, grade repetition, grade and year fixed effects, and controls for teacher experience:[27]

$$\begin{aligned} y_{ijgst} &= f_g(y_{i_{ij'g-1s't}}) + x'_{ijgst}\beta + h(exp_{jt}) + \psi_{ijgst}, \text{ where} \\ \psi_{ijgst} &= \mu_j + \theta_{jt} + \epsilon_{ijgst}, \text{ and } \mu_j = a_j + \underline{e_j}. \end{aligned} \tag{17}$$

The EB estimator uses several years of data for each teacher to construct an optimally-weighted average of classroom-level residual test scores in order to separate teacher ability, μ_j, from classroom-specific shocks, θ_{jt}, and student-level noise, ϵ_{ijgst}. While we implement the same procedure as previous studies, we note that in our setting, EB estimates consist of both incentive-invariant ability and baseline ABCs effort, $\hat{\mu}_j = \widehat{(a_j + \underline{e_j})}$.

3. NCLB-Induced Effort Response: We define a student as 'marginal' if she is predicted to score within $+/-4$ developmental scale points of the proficiency cutoff, noting that various alternative definitions yield the same outcome. For each classroom, the relative incentive strength measure, m_{jt}, is defined as the fraction of students in that classroom who are marginal. We then identify the component of teacher-year quality that is attributable to NCLB effort incentives by regressing teacher-year fixed effects $\hat{q}_{jt}$ on m_{jt}, while holding

[24] For estimators that allow teacher ability to 'drift' over time, see Goldhaber and Hansen (2013), Chetty *et al.* (2014a), Rothstein (2014), and Bacher-Hicks *et al.* (2014).

[25] Assuming that baseline effort is fixed over time in our setting is reasonable, given that the pre-existing value-added incentive scheme is approximately uniform in its effects on teacher effort (see Macartney *et al.* (2015)).

[26] Due to the differential timing of the 2001 math developmental scale change in third grade, and the unavailability of second grade scores in 1996, we run a separate regression for third grade from 1998 to 2000.

[27] We paramaterize the experience function by including indicators for each level of experience from zero to five years, with the omitted category being teachers with six or more years of experience. We choose this specification to be consistent with Chetty *et al.* (2014a). Wiswall (2013) shows that such a restrictive choice may be restrictive, biasing estimates of the dispersion in teacher quality – an issue we intend address in future work.

constant teacher incentive-invariant ability (and baseline effort) and teacher experience:

$$\hat{q}_{jt} = \alpha + \rho m_{jt} + \lambda(\widehat{a_j + \underline{e_j}}) \ + w(exp_{jt}). \tag{18}$$

Once accountability pressure due NCLB is introduced (captured by m_{jt}), teachers may exert additional effort, according to the amount of pressure they face. We thus test whether there is a systematic relationship between $\hat{q}_{jt}$ and m_{jt} due to NCLB in 2003 but no relationship prior. We identify effort as a predicted value from equation (18), $e(m_{j2003})$, representing the response to the new incentive scheme and capturing the relationship (conditional on ability and experience) between teacher-year performance and the classroom fraction of marginal students in 2003, given by:

$$e(m_{j2003}) = \hat{\rho} m_{j2003}. \tag{19}$$

V.B. Results from the Ability/Effort Decomposition

This subsection presents the results of the ability and effort decomposition in three parts. First, we present summary statistics for the key teacher-level variables of interest and report our main estimates for the effects teacher ability and teacher effort. Second, we supplement these estimates by relying entirely on within-teacher variation in performance and showing that greater improvement occurs among teachers with stronger incentives. Third, we assess the main threat to identification, showing that students were not sorted differentially to teachers (based on teacher ability) in 2003 in a way that can cause bias in our results.

V.B.1 Identifying Ability and Effort

Table 2 reports means and standard deviations of the main teacher-level variables and estimates of interest, while Figure 3 presents the incentive-invariant teacher ability distribution,[28] where incentive-invariant ability is defined as the EB estimate from equation (17). The distributions are similar among third, fourth, and fifth grade teachers, with means of -0.07, -0.09 and -0.06 developmental scale points, and standard deviations of 1.68, 1.37 and 1.3 developmental scale points, respectively. The EB estimator shrinks teacher effects to-

[28]This includes the baseline effort level discussed in the prior section, which is also assumed invariant to NCLB incentives.

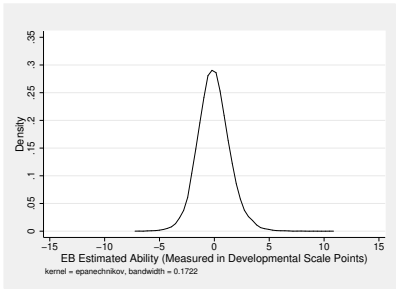
ward the mean to minimize the influence of measurement error, thus resulting in observed standard deviations that are downward-biased. The estimated 'true' standard deviations reported in Table 2 are slightly larger, at 2.16, 1.63 and 1.63 developmental scale points. Averaged across grades, mean teacher ability is -0.074 scale points and its standard deviation is 1.79 scale points, or equivalently 0.18 student-level standard deviations. (The latter is within the upper bound of the range found by most previous work that estimates teacher quality.)

Table 2: Teacher Performance Variables

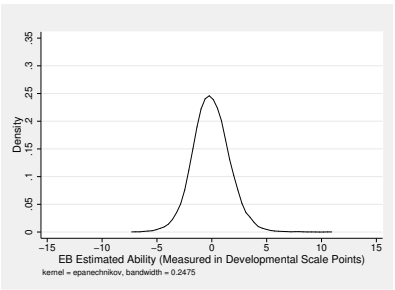
	Teacher-Year Value-Added			Estimated Ability			Fraction of Marginal Students, m_{jt}			Estimated Effort		
	Gr. 3	Gr. 4	Gr. 5	Gr. 3	Gr. 4	Gr. 5	Gr. 3	Gr. 4	Gr. 5	Gr. 3	Gr. 4	Gr. 5
Mean	-0.17	-0.11	0.19	-0.07	-0.09	-0.06	0.33	0.21	0.23	0.56	0.80	0.45
Observed SD	2.65	2.80	2.31	1.68	1.38	1.30	0.16	0.14	0.15	0.27	0.64	0.36
Estimated SD	-	-	-	2.16	1.63	1.63	-	-	-	-	-	-
Obs.	24,105	22,246	20,596	6,547	7,816	7,046	17,371	16,075	14,817	2,144	2,598	2,570

Notes: This table presents means and standard deviations of the main teacher-level variables of interest. Summary statistics for teacher-year VA measures and for the fraction of marginal students in classrooms are calculated over all available teacher-year observations from 1997 to 2003. Due to the unavailability of second grade scores in 1996 and the change to the math developmental scale in 2001, we are unable to calculate marginal status for third graders in 1997 and 2001, and for fourth and fifth graders in 2002, thus explaining the smaller sample sizes. Summary statistics for estimated ability are calculated over all teacher-grade observations, where we include a teacher in a grade-specific distribution if she is ever observed teaching in that grade. A given teacher can be in more than one grade-specific distribution. The observed standard deviation is the raw standard deviation, while the estimated standard deviation is the estimate of the true standard deviation of teacher ability, obtained from the EB procedure. Summary statistics for estimated teacher effort are calculated across all teacher observations in 2003.

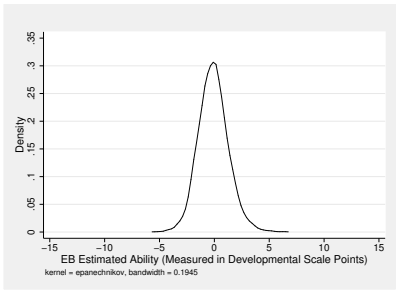
Panel (a) of Table 3 reports the underlying estimates of ρ from equation (18), while the top panels of Figure 4, (a) through (c), show the grade-specific partial relationships between $\hat{q}_{jt}$ and m_{jt}. For each grade, we plot the relationships that prevail in 2003 and a pooled regression of all pre-NCLB years that additionally includes year fixed effects. In 2003, there is a clear increasing relationship between the part of the teacher-year effect unexplained by ability and experience and the proportion of marginal students in the classroom. Relative to the 2003 raw-data patterns found in Figure 2, we add plots of the corresponding relationships in pre-NCLB years, noting that our three-part procedure accounts for any correlation between incentive-invariant ability and NCLB incentives. The estimates in panel (a) of Table 3 imply that, conditional on teacher ability and experience, a one standard deviation increase in the proportion of marginal students within a classroom is associated with a 9



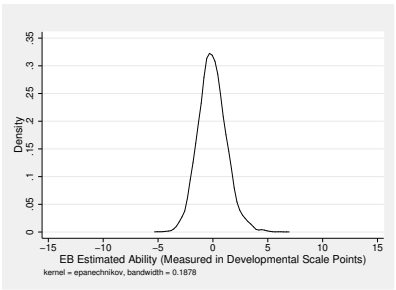
(a) All Grades



(b) Grade Three



(c) Grade Four



(d) Grade Five

Notes: This figure shows the distributions of teachers' incentive-invariant abilities (which include base level effort). To construct the figures, we estimate equation (17), and construct EB estimates of teacher ability. Panel (a) shows the distribution of ability across all teachers. Panels (b), (c), and (d) show the distributions for teachers in third, fourth and fifth grades, respectively. We include a teacher in a grade-specific distribution if she is ever observed teaching in that grade. A given teacher can be in more than one grade-specific distribution.

Figure 3: Incentive-Invariant Ability Distributions

percent, 22 percent, and 16 percent standard deviation increase in teacher-year VA in third, fourth, and fifth grade, respectively. As expected, there is virtually no relationship in the pre-NCLB years, once we account for the negative correlation between teacher ability and the classroom fraction of marginal students.[29]

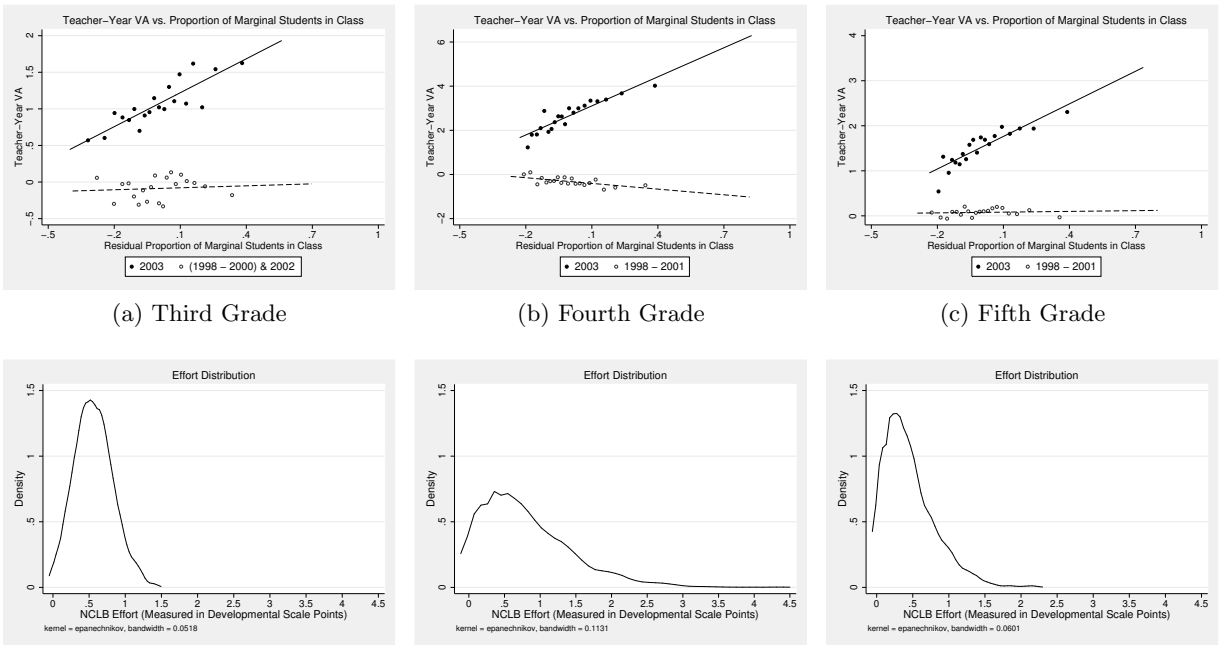
Incentives under NCLB cause modest *average* improvements in student test scores: a one standard deviation increase in the fraction of marginal students in a classroom corresponds to test score improvements of 0.03, 0.06, and 0.04 student-level standard deviations in third, fourth, and fifth grade, respectively. Panels (d) through (f) of Figure 4 present the full distributions of effort in each grade in 2003, where effort is constructed as the fitted value from equation (19). The last three columns in Table 2 report the associated means and standard deviations. Mean effort for third, fourth and fifth grades is 0.56, 0.8, and 0.45 developmental scale points, respectively (the average across grades is 0.61 points). Dispersion in effort across grades is 0.27, 0.64 and 0.36 points, and is 0.48 scale points across all grades, which

[29]When controlling for teacher ability in the pre-NCLB period, we apply the jack-knife EB estimator, which uses information from all *other* years excepting the one in question (Chetty *et al.*, 2011). This avoids mechanical correlation in measurement error driving any of the results.

Table 3: The Effects of NCLB Incentives on Teacher Performance

	Panel (a): Teacher-Year VA as Dependent Variable					
	Third Grade		Fourth Grade		Fifth Grade	
	2003	Pre-NCLB	2003	Pre-NCLB	2003	Pre-NCLB
Effect of m_{jt}	1.55***	0.09	4.39***	-0.85***	2.41***	0.06
	(0.20)	(0.13)	(0.32)	(0.15)	(0.23)	(0.16)
Observations	2,144	10,452	2,598	11,551	2,570	10,609
	Panel (b): Change in Teacher-Year VA as Dependent Variable					
	Third Grade		Fourth Grade		Fifth Grade	
	2003	2000	2003	2000	2003	2000
Effect of m_{jt}	2.08***	-0.40	4.11***	-0.64**	2.48***	-0.53
	(0.17)	(0.27)	(0.32)	(0.30)	(0.24)	(0.35)
Observations	2,651	2,393	2,453	2,385	2,397	2,187

Notes: In panel (a), we present estimates of ρ from grade-specific regressions of equation (18). In the year 2003 regression, additional controls include teacher ability and teacher experience. The result in the pre-NCLB columns comes from a pooled regression of all pre-NCLB years that additionally includes year fixed effects. For third grade, the pre-NCLB years stretch from 1998 to 2000, and 2002; for fourth and fifth grade, they stretch from 1997 to 2001. In panel (b), we regress the change in teacher-year VA from 2002 to 2003 or from 1999 to 2000 on the fraction of marginal students in classrooms in 2003 and 2000, respectively, as well as cubic functions of 2002 and 1999 teacher-year VA. The reported coefficients are the effects of the fraction of marginal students within classrooms. Standard errors clustered at the school level appear in parentheses. *** denotes significance at the 1% level;** denotes significance at the 5% level; and * denotes significance at the 10% level.



(a) Third Grade (b) Fourth Grade (c) Fifth Grade

(d) Third Grade (e) Fourth Grade (f) Fifth Grade

Notes:This figure illustrates teachers' 2003 effort responses. In panels (a) to (c), we present grade-specific partial relationships between teacher-year effects and the fraction of students in a teacher's class who were marginal. To construct these figures, we first residualize m_{jt} with respect to the other controls in equation (18). For the pre-NCLB years, these controls also include year fixed effects. The horizontal axis measures residualized m_{jt}. We group teacher-year observations in 20 equal-sized groups (vingtiles) of the residualized m_{jt} distribution on the horizontal axis. Within each bin, we calculate the average residualized m_{jt} and the average teacher-year effect. The circles in each panel represent these averages. The lines represent the associated linear effects, estimated on the underlying teacher-year data. In panels (g) to (i), we present grade-specific densities of 2003 effort levels. To construct these figures, we first obtain 2003 effort for each teacher by taking the linear prediction (fitted value) from equation (19). We then plot the distributions of these effort levels separately by grade.

Figure 4: Effort Predictions and Effort Distributions in 2003

amounts to 0.05 student-level standard deviations. Taken together, the evidence indicates that there is meaningful variation in teachers' effort in 2003, strongly correlated with the fraction of students in their classes who are marginal.

V.B.2 Within-Teacher Performance Improvements

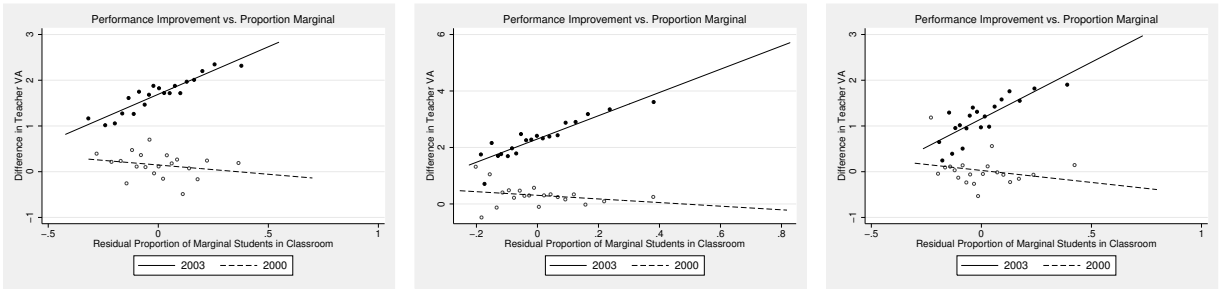
We show the further relevance of NCLB incentives by documenting that they cause within-teacher performance improvements. Specifically, we construct the difference between 2003 and 2002 teacher-year fixed effects, as

$$\hat{q}_{j2003} - \hat{q}_{j2002} = a_j + e_{j2003} + \bar{\epsilon}_{j2003} - (a_j + e_{j2002} + \bar{\epsilon}_{j2002}) = e_{j2003} - e_{j2002} + \bar{\epsilon}_{j2003} - \bar{\epsilon}_{j2002}, \quad (20)$$

and estimate the relationship between this difference and the fraction of marginal students whom that teachers faced in 2003, m_{j2003}. To ensure that mean reversion is not responsible

for the results, we also control for a cubic function of 2002 teacher-year VA.

Panels (a) through (c) of Figure 5 show the partial relationships between performance improvement in 2003 and m_{j2003}, while panel (b) of Table 3 reports the underlying slope coefficients. As expected, within-teacher performance improvements are clearly increasing in the fraction of marginal students within classrooms in 2003. The transition from 1999 to 2000 is used as a placebo control in each grade, revealing a flat relationship and lending support to the claim that the 2003 patterns reflect teachers improving performance as a result of NCLB effort incentives.



(a) Third Grade (b) Fourth Grade (c) Fifth Grade

Notes:This figure illustrates teachers' 2003 effort responses. In panels (a) to (c), we depict the relationship between the change in teachers' annual performance from 2002 to 2003 and from 1999 to 2000 and the fraction of students in their classes who were marginal in 2003 and 2000, respectively. To construct the panels, we first construct the change from $t-1$ to t between each teacher's teacher-year fixed effects from those years, as shown in equation (20). For each teacher-year, we then calculate the fraction of students in the 2000 or 2003 class who were marginal. We group teacher-year observations into 20 equally-size (vingtiles) bins of the fraction marginal distribution on the horizontal axis. Within each bin, we calculate the average proportion of marginal students and the average change in teacher-year fixed effects. The circles in each panel represent these averages. The lines represent the associated linear fits, estimated on the underlying teacher-year data.

Figure 5: Within-Teacher Performance Improvements

V.B.3 Rival Hypothesis: Addressing Potentially Differential Sorting of Students to Teachers

To infer teacher effort, we compare the relationship that prevailed between a teacher's performance and the fraction of students in her class who are marginal in 2003 with the relationship that prevailed in previous years. Since we see a positive relationship in 2003 and no relationship in prior years, we argue that the 2003 relationships reflect changes in teachers' effort, drawing on the theoretical predictions concerning responses to proficiency-count systems (see Proposition 1).

A competing explanation is that students were sorted *differentially* to teachers in 2003 such that high (incentive-invariant) ability teachers received larger fractions of marginal students. While we control for teacher ability in the process of estimating effort responses, if high-ability teachers were better able to respond to the demands of NCLB, we might worry

that this non-linear relationship between teacher and student ability is driving the results rather than additional effort being exerted by a *given* teacher.

A natural way to evaluate this rival hypothesis is to test whether the relationship between the fraction of marginal students in a classroom and teacher ability changes in 2003. We conduct this test by regressing the fraction of marginal students in each class on grade and year fixed effects, our measure of combined teacher incentive-invariant ability and baseline effort, and an interaction of that term with a year-2003 indicator:

$$m_{jt} = \alpha_0 + \lambda_g + \lambda_t + \beta_1(\widehat{a_j + \underline{e}_j}) + \beta_2(\widehat{a_j + \underline{e}_j}) \times 1(t = 2003) + \epsilon_{jt}, \ \forall \ t \leq 2003. \qquad (21)$$

If principals began differentially sorting students to teachers on the basis of ability in 2003, we would expect to find the coefficient on ability and baseline effort not equal to zero ($\beta_2 \neq 0$).

Table 4: Tests for Differential Sorting of Students to Teachers in 2003

	(1) Full Sample	(2) Third Grade	(3) Fourth Grade	(4) Fifth Grade
Ability	-0.0034***	-0.0010	-0.0046***	-0.0055***
	(0.0008)	(0.0010)	(0.0011)	(0.0017)
$1(t = 2003) \times$Ability	-0.0033**	-0.005***	-0.0045**	0.0027
	(0.0014)	(0.0019)	(0.0019)	(0.0025)
N	39,932	12,599	14,151	13,182

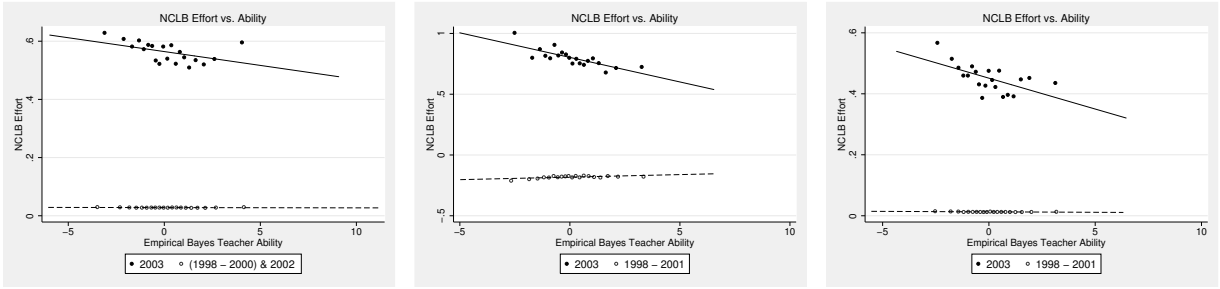
Notes: This table presents the results of regressions based on equation (21). The dependent variable in each column is the fraction of students in a teacher's class who are marginal. Teacher ability is estimated using the EB estimator from equation (17), and we use the leave-year-out (or jack-knife) EB estimate in pre-NCLB years to avoid mechanical correlation between EB estimates and outcomes. Standard errors clustered at the school-level appear in parentheses. *** denotes significance at the 1% level;** denotes significance at the 5% level; and * denotes significance at the 10% level.

Table 4 shows the results from estimating variants of equation (21). Overall, there is a small negative relationship between the fraction of marginal students who are in a teacher's class and that teacher's incentive-invariant ability. This reflects the relatively low test score proficiency standard in North Carolina and the sorting of historically low-performing students to low-ability teachers. The estimates in column (1) imply that a one standard deviation better than average teacher has 0.61 percentage points fewer marginal students in her class (which corresponds to a 2.3 percent reduction relative to the mean fraction). The sorting patterns appear to change slightly in 2003, but such that high-ability teachers receive *smaller* fractions of marginal students than in the pre-NCLB period. This change is in the opposite

direction to that required to bias our results upward.

While we are able to isolate the relationship between teacher performance and NCLB incentives *conditional* on teacher ability, the low test score proficiency target and the sorting patterns of students to teachers result in slightly stronger effort incentives for low-ability teachers. Figure 6 shows grade-specific relationships between $e(m_{jt})$ and $a_j + \underline{e}_j$. For each grade, we plot the relationships that prevail in 2003 and the pooled pre-NCLB control years. In 2003, there is a clear decreasing relationship between NCLB effort and incentive-invariant ability. In the control years, the estimated functions are virtually flat.

The lack of a relationship between 'placebo' NCLB effort and teacher ability is reassuring, as NCLB incentives did not operating in those years: $e(m_{jt})$ is identified by holding ability constant, implying that we automatically account for the correlation between ability and m_{jt} in the estimation routine. Since m_{jt} is not associated with any incentives prior to NCLB, there is no remaining relationship to identify. In contrast, in 2003, m_{jt} reflects NCLB incentives, and the analysis reveals stronger effort responses among the teachers who faced stronger incentives. The slope coefficients are all significant at the one percent level and are -0.01, -0.04, and -0.02, for third, fourth and fifth grade, respectively. Thus, a one standard deviation *lower* ability teacher in 2003 exerted approximately 0.02, 0.07, and 0.03 developmental scale points worth of additional effort in third, fourth grade, and fifth grade. These differences are very small, however, corresponding to 0.002, 0.007, and 0.003 student-level standard deviations.



(a) Grade Three (b) Grade Four (c) Grade Five

Notes: This figure illustrates the relationship between NCLB teacher effort, obtained as the fitted value from equation (19), and teacher incentive-invariant ability. We construct the figure by first grouping teachers into 20 equal-sized bins (vingtiles) of the ability distribution. Within each bin, we calculate average ability and average NCLB effort. The circles in each panel represent these averages. The lines represent the associated linear effects, estimated using the underlying teacher-year data.

Figure 6: The Relationship Between Teacher Effort and Incentive-Invariant Ability

VI. Estimating the Persistence of Teacher Ability and Effort

In this section, we address whether teacher ability and effort persist at different rates. These will prove to be vital ingredients in the policy analysis that follows. As in the previous section, we outline our estimation strategy before presenting the empirical results.

VI.A. Estimating the Persistence of Ability

We begin by estimating the persistence of ability and baseline effort in a reduced-form way, following the previous literature (for example, Chetty *et al.*, 2014b). Specifically, we regress student test scores in period $t+n$ on the full control vector from the Empirical Bayes regression (equation (17)) and the ability of the teacher who taught the students in period t:

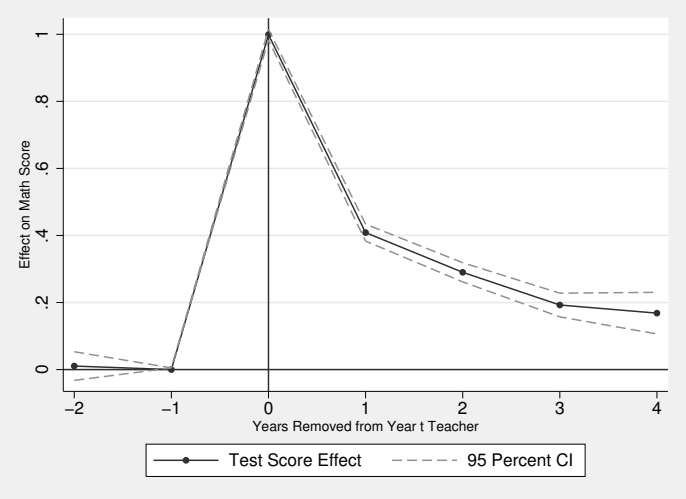
$$y_{ijgst+n} = f_g(y_{i_{ij'g-1s't}}) + x'_{ijgst}\beta + h(exp_{jt}) + \beta_n(\widehat{a+\underline{e}})_{jt} + \epsilon_{ijgst}. \tag{22}$$

Here, ability is measured with a jack-knife EB estimator, which uses information from all years except the current year to form an estimate of teacher ability. This prevents mechanical correlation between measurement error in test scores and in teacher ability from confounding the results (Chetty *et al.* 2014a). The coefficient β_n represents the degree to which the effect of teacher ability from year t persists forward to influence test scores in year $t+n$.

Figure 7 presents the estimated β_n coefficients from regressions that include test scores exclusively from the pre-NCLB period. As a check, teachers do not affect their students test scores in the years before they are matched with these students, as shown by the estimate at $t-2$.[30] As expected, a one unit better-than-average teacher in year t improves student test scores by 1 developmental scale point, on average. The contemporaneous effect of teacher ability fades away overtime, as 41 percent of the initial effect persists to affect test scores in period $t+1$, and only 20 percent remains by period $t+4$.

These results align closely with the prior literature (Jacob, Lefgren, and Sims, 2010; Chetty *et al.* 2014a). Having estimated the one-year ahead persistence of ability at 0.41, we seek to obtain estimates of the persistence of teacher effort and contrast the two per-

[30]Since we control for once-lagged test scores when estimating teacher ability, the coefficient at $t-1$ is identically zero.



Notes: This figure reports estimates of the β_n coefficients from equation (22). Each estimate is obtained from a separate regression in the pre-NCLB period from 1997 to 2002. The horizontal axis measures the number of year students are removed from their period-t teacher while the vertical axis measures the impact of the period-t teacher on students' test scores in period $t+n$. The dark circles represent the estimated effects while the dashed lines represent the 95 percent confidence intervals with the associated standard errors clustered at the school level.

Figure 7: Persistence of Teacher Ability and Baseline Effort in Pre-NCLB Period

sistence effects. Estimating the persistence of effort is a more challenging exercise, as the strong within-student correlation of NCLB effort over time implies that one must account for contemporaneous effort to avoid overstating the persistent effects of lagged effort. Yet, because effort persistence helps determine student predicted scores, and educators make effort decisions based on anticipated student performance, contemporaneous effort is itself a function of the persistence parameter. We also must take into account potential changes to ABCs effort to avoid confounding school-level ABCs-related improvement with student-level effort persistence.

We now develop a strategy for estimating the persistence of effort and discuss the identification issues in greater detail.

VI.B. Estimating the Persistence of Effort

VI.B.1 Measuring Student-Level Effort

In the analysis above, we identified teacher effort at the classroom level. While natural when assessing whether teacher VA depends on performance incentives in a systematic way, given that teacher-year VA itself represents classroom-level residual test score gains, classroom effort is not our preferred measure for estimating persistence. Instead, more variation in the

data can be exploited by constructing a student-level measure, and then investigating the rate at which student-specific effort persists.[31]

We measure student-level effort by building on the non-parametric patterns in Figure 1, which show that the introduction of NCLB had pronounced non-linear effects on student test scores, in a way that is consistent with strong teacher effort responses to the scheme.[32] In particular, students near the passing threshold scored approximately 3 developmental scale points higher than their predicted scores, a gain equivalent to 30 percent of the test score standard deviation.

The incentive strength measure on the horizontal axis is constructed in three steps. In the first step, we predict student performance in a flexible way in pre-NCLB years using several covariates, including lagged test scores.[33] In the second step, we use the saved regression coefficients from the first step to construct a predicted score for each student in 2003, denoted $\hat{y}_{ijgs2003}$, by combining those coefficients with updated (but pre-determined) student covariates. In the third step, we use the NCLB target, $y_g^{T,nclb}$, to compute incentive strength as the difference between the predicted score and the target, written $\pi_i = \hat{y}_{ijgs2003} - y_g^{T,nclb}$.[34]

The difference $\hat{y}_{ijgs2003} - y_g^{T,nclb}$ captures how 'far away' each student is from reaching proficiency status without additional teacher effort. In terms of the model in Section IV, we write the predicted score $\hat{y}_{ijgst}$ as the sum of all non-NCLB effort inputs in the production technology, $\hat{y}_{ijgst} = \gamma y_{ij'g-1s't-1} + a_j + e_s^{*abcs}$. This represents the test score each student would earn in 2003 had NCLB not been introduced. Taking the difference between realized and predicted scores in 2003 allows us to write the variable measured on the vertical axis in Figure 1 as a function of NCLB effort and random shocks to test scores: $y_{ijgst} - \hat{y}_{ijgst} = e_{ijgst}^{nclb} + \epsilon_{ijgst}$. Since effort incentives are strongest for students who are predicted to score near the proficiency threshold ($\hat{y}_{ijgst} - y_g^{T,nclb} \approx 0$), these students should receive the most

[31] It is well-established that much of the variation in student test scores occurs within rather than across schools (Kane and Staiger, 2002). Thus, much of the variation in students' predicted test scores is likely to occur within schools. In turn, since NCLB proficiency targets are *fixed* across students, and the distance between predicted scores and the target determines student-level incentive strength, much of the variation in educators' effort incentives is also likely to occur within schools. We therefore prefer to use a student-level measure of effort in order to make use of all of the available variation.

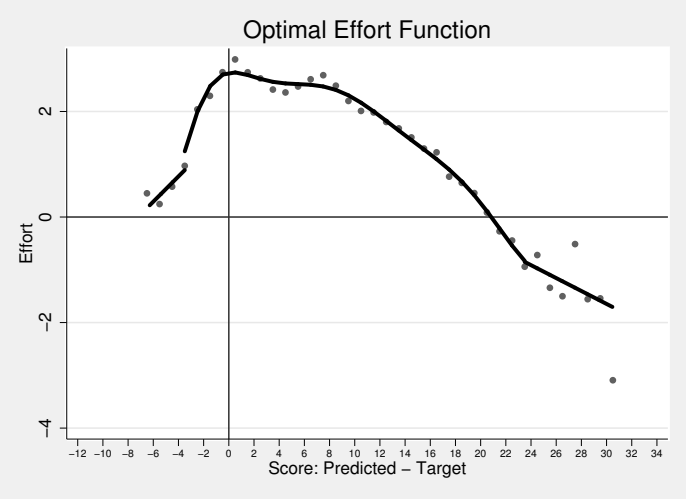
[32] More details are available in Macartney *et al.* (2015).

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

[34] The predicted score is invariant to any changes occurring in 2003, with variation in incentive strength arising from the proficiency target $y_g^{T,nclb}$ becoming relevant under NCLB.

additional effort and gains over predicted scores should highest for them, exactly as shown in the figure.[35]

These non-parametric patterns are used to estimate a student-specific effort function that takes $\hat{y}_{ijgst} - y_g^{T,nclb}$ as its argument. We estimate the function by first differencing the year 2003 and 2000 profiles in Figure 1 and then fitting a tenth-order polynomial to the binned data using a weighted regression, with the weights capturing the number of students in each bin. The resulting effort function, denoted by $e^{nclb}(\hat{y}_{igst} - y_g^{T,nclb})$, is plotted in Figure 8. The value of $e^{nclb}(\hat{y}_{igst} - y_g^{T,nclb})$ is used as the student-specific value of effort each student receives in 2003.



Notes: This is Figure 5(b) from Macartney *et al.* (2015). The horizontal axis is the same as in Figure 1. To construct this figure, we first take the bin-specific differences between the year 2003 and the year 2000 vertical-axis variable in Figure 1. The dots represent the resulting within-bin differences. We then estimate a tenth-order polynomial on the binned-data, weighting the regression by the number of student observations (across both 2003 and 2000) within each bin. At the extremes of the horizontal axis (less then -3 and greater than 22), we estimate linear regressions. The lines trace out the empirical effort function that results from these regressions.

Figure 8: Student-Specific Effort Function $e^{nclb}(\hat{y}_{igst} - y_g^{T,nclb})$

VI.B.2 Identifying the Persistence of Effort

Our measure of student-level effort, e_{ijgst}^{nclb}, indicates the effort received by student i, who is in the classroom of teacher j in grade g at school s in time t. With this measure in hand, we turn to estimating the rate at which effort persists. In equation (13) of the model in

[35]To ensure that we do not systematically under- or over-predict test scores for certain parts of the distribution, we conduct the same exercise with a pre-reform year. As can be seen in Figure 1, we do a good job predicting test score outcomes in 2000 throughout the distribution, lending credence to the claim that the 2003 patterns reflect student-specific NCLB effort.

Section IV, we posit that test scores in 2004 are determined by the persistent effect of the non-NCLB effort component of test scores ($y_{ij'g-1s't-1} - e^{nclb}_{ij'g-1s't-1}$), the contemporaneous effects of teacher ability, ABCs-related effort, and NCLB-related effort, the persistent effect of prior NCLB effort, and a random shock to test scores:

$$y_{ijgst} = \gamma(y_{ij'g-1s't-1} - e^{nclb}_{ij'g-1s't-1}) + a_j + \tilde{e}^{abcs}_s + e^{nclb}_{ijgst} + \gamma^e e^{nclb}_{ij'g-1s't-1} + \epsilon_{ijgst}. \tag{23}$$

Here, $\tilde{e}^{abcs}_s$ denotes the school-specific ABCs effort that results after NCLB's disruption of ABCs targets.

When modelling predictions about students' likely performance without additional NCLB effort in 2004, we assume educators know the persistence rate of effort, γ^e, forming the prediction as the sum of (i) the persistence of non-NCLB effort inputs, (ii) the effects of teacher ability and baseline ABCs effort, and (iii) the persistence of NCLB effort. Drawing a distinction between the component of the predicted score that is independent of the persistence of NCLB effort and the part that depends on it, we let $\tilde{y}_{ijgst}$ denote the test score students would earn in 2004 had NCLB not been enacted in the prior year. In that case, there is no contemporaneous or persistent effect of NCLB effort and ABCs effort remains in steady state, implying that $\tilde{y}_{ijgst} \equiv \gamma(y_{ij'g-1s't-1} - e^{nclb}_{ij'g-1s't-1}) + a_j + e^{*abcs}_s$. Substituting $\tilde{y}_{ijgst}$ into the production technology in equation (23), test scores in 2004 are expressed as the sum of predicted scores and contemporaneous effort responses:

$$y_{ijgst} = \underbrace{\tilde{y}_{ijgst} + \gamma^e e^{nclb}_{ij'g-1s't-1}}_{\text{score predicted by educators}} + \underbrace{e^{nclb}_{ijgst} + \hat{e}^{abcs}_{st}}_{\text{effort responses}} + \epsilon_{ijgst}, \tag{24}$$

where $\hat{e}^{abcs}_s = \tilde{e}^{abcs}_{st} - e^{*abcs}_s$ denotes the 2004 school-specific deviation from steady-state ABCs effort caused by NCLB's disruption to ABCs targets.

When making decisions about how much NCLB-related effort to devote to each student in 2004, schools again take into account the distance between students' predicted scores and the NCLB proficiency target, responding based on the NCLB effort function:

$$e^{nclb}(\tilde{y}_{igst} + \gamma^e e^{nclb}_{ij'g-1s't-1} - y^{T,nclb}_g). \tag{25}$$

Equation (25) shows that the effort decision in 2004 depends on the effort students received

in 2003 and the parameter γ^e, thus highlighting the correlation of effort over time.[36]

We account for this correlation using the following estimating equation for 2004 test scores, motivated by equation (24):

$$\begin{aligned} y_{igs2004} - \tilde{y}_{igs2004} =& \gamma^e e^{nclb}(\hat{y}_{ig-1s'2003} - y_{g-1}^{T,nclb}) + \theta e^{nclb}(\tilde{y}_{igs2004} + \gamma^e e_{ig-1s'2003}^{nclb} - y_g^{T,nclb}) \\ &+ \rho \hat{e}_{s2004}^{abcs} + \epsilon_{igs2004}. \end{aligned} \tag{26}$$

Taking the components of this equation in turn, the dependent variable is the difference between the realized test score and the score students would have earned in 2004 had NCLB not taken effect in 2003. By definition, it captures the effects of all new test score determinants that result from the incentives created by NCLB, which include the persistent effects of NCLB effort in 2003 and the contemporaneous effects of both NCLB and ABCs effort.

We define an empirical analogue to $\tilde{y}_{ijgst}$ (the test score students would earn in 2004 had NCLB not been enacted in the prior year) as the 'counterfactual' predicted score, which is constructed for students in grade g in 2004 by using our test-score prediction equation for that grade (estimated prior to NCLB) and substituting *predicted* grade $g-1$ test scores from 2003 for realized 2003 test scores. Realized scores contain NCLB effort from 2003 and are thus a function of γ^e. Using predicted scores in place of realized scores then ensures that the counterfactual predicted score represents the score students would earn in 2004 if there were no NCLB incentives in 2003.

Since the effort function *and* its inputs are both known in 2003, the effort devoted to each student in 2003 is given by $e^{nclb}(\hat{y}_{ig-1s'2003} - y_{g-1}^{T,nclb})$ and is thus observed to the econometrician. In 2004, we let $e^{nclb}(\tilde{y}_{igs2004} + \gamma^e e_{ig-1s'2003}^{nclb} - y_g^{T,nclb})$ represent the unknown value of effort and allow the parameter θ to amplify or diminish the NCLB effort response one year after the introduction of the reform.

[36] Our structural assumption is that teachers choose their effort levels according to the underlying effort function. Throughout the estimation routine, we allow effort decisions in future years to be amplified or diminished relative to the first year of NCLB, but continue to assume that teacher effort is primarily determined by the estimated effort function $e^{nclb}(\cdot)$. Under this assumption, teachers can exert more or less overall effort in response to NCLB over time but always choose to exert relatively more effort toward students who are predicted to score on the margin of the passing threshold.

Maximum Likelihood Routine

Since the input to the 2004 effort function depends on the persistence rate γ^e, we are unable to observe the level of effort received by students in 2004 directly without knowing the persistence rate. Yet in order to credibly identify the persistence rate, we need to account for the correlation of effort across time. We therefore use a maximum likelihood routine to jointly estimate γ^e, effort in 2004, $e^{nclb}(\tilde{y}_{igs2004} + \gamma^e e^{nclb}_{ig-1s'2003} - y_g^{T,nclb})$, and the scale factor modifying the 2004 effort function, θ.

Our routine also accounts for potential changes to ABCs effort incentives in 2004. The conceptual discussion in Section IV motivated the idea that the magnitude of NCLB's disruption to ABCs effort should in large part be a function of the difference between the persistence rates of non-effort and effort inputs and the average level of NCLB effort received by students in a given school in 2003, $(\gamma^a - \gamma^e) \sum_{g=3}^{G_s} \bar{e}^{nclb}_{g-1st-1}$. We use the student-level effort function $e^{nclb}(\hat{y}_{igst} - y_g^{T,nclb})$ to calculate average school-level effort from 2003, which allows us to control for it directly in estimation routine in order hold constant the effects of ABCs incentives in 2004. Equation (26) is therefore modified by setting $\hat{e}^{abcs}_{s2004} = \bar{e}^{nclb}_{s2003}$, with ρ governing the effect of average effort (across all students) from 2003 on test scores in 2004.[37]

We implement the maximum likelihood routine by assuming $\epsilon_{igs2004} \sim N(\mu, \sigma^2)$, treating μ and σ^2 as additional parameters to be estimated. The full parameter vector is given by $\omega = [\gamma^e, \theta, \rho, \mu, \sigma^2]'$ and the main parameters of interest, γ^e, θ, and ρ, are all separately identified. In what follows, we establish separate identification by arguing that γ^e and θ are separately identified, first by ignoring ABCs effort incentives, then expanding the argument to include ABCs incentives, showing further that ρ is separately identified from both γ^e and θ.

Identification Argument without ABCs Incentives

Conceptually, separate identification of γ^e and θ requires that conditional on 2003 effort, $e^{nclb}(\hat{y}_{ig-1s'2003} - y_{g-1}^{T,nclb})$, there is remaining variation in 2004 effort, $e^{nclb}(\tilde{y}_{ig2004} + \hat{\gamma}^e e^{nclb}_{ig-1s'2003} - y_g^{T,nclb})$, and vice-versa.[38] Such variation is guaranteed by the non-linear shape of the effort

[37] We calculate $\bar{e}^{nclb}_{s2003}$ as a jack-knife mean, leaving out the effort received by student i, to ensure that the estimates of γ^e and ρ are not confounded.

[38] This condition is different from requiring that, conditional on the *argument* of the effort function in 2003, $\hat{y}_{ig-1s'2003} - y_{g-1}^{T,nclb}$, there is remaining variation in the *argument* of the effort function in 2004, $\tilde{y}_{ig2004} + \gamma^e e^{nclb}_{ig-1s'2003} - y_g^{T,nclb}$. In general, it is not possible to estimate the effects of *incentive strength* separately from different time periods, as the definition of the counterfactual predicted score and the linear form of NCLB incentives imply that student-

function in 2003, which ensures that two students with the same level of effort in 2003 can have different levels of NCLB incentive strength and, correspondingly, different levels of 2004 NCLB effort.

For example, referring to Figure 8, consider two students who each have a 2003 effort level of 2 developmental scale points, but one student has an incentive strength value of -3 scale points while the other has a value of 1 scale point. Despite having the same level of effort in 2003, the students have different predicted scores in 2003 and continue to have different predicted scores in 2004. The student with prior incentive strength of -3 is still predicted to score relatively poorly but the extra effort she received last year moves her up in the incentive strength distribution, making her more marginal with respect to the test score target. This student receives *more* effort in 2004 than she did in 2003. The student with prior incentive strength of 11 also gets a bump in his predicted score, which also moves him up in the incentive strength distribution, but to a point where he receives *less* effort in 2004 because he is even further away from the proficiency threshold and therefore less marginal.

Following similar logic, we use a more general argument to establish that, conditional on the 2003 effort function, identification of the parameters of interest requires a minimal assumption about the form of the 2004 effort function, namely that it should not be flat. We label this as the requirement that the function be "non-uniform." Although we assume the same functional form for the effort function in 2004 as in 2003, this assumption is not required for identification. Indeed, any non-uniform effort function in 2004 would be sufficient.

To see this, suppose that the 2004 effort response is determined by some arbitrary non-uniform function and – as above – consider any two inframarginal students (one with a predicted score below the proficiency target and one above it) who receive the same level of effort in 2003, due to the non-monotonic nature of the effort function. Incentive strength in 2004 shifts rightward for each student by the common amount of 2003 effort that persists. A non-uniform effort function in 2004 then guarantees that at least some student pair satisfying the identical-effort condition in 2003 receives divergent levels of effort in 2004. Indeed, there is zero variation in effort within all such student pairs only if the 2004 effort function is uniform (or flat), implying that any non-monotonic effort function in 2003 and non-uniform function in 2004 are sufficient for identification.

specific incentive strength in 2004 is nearly a perfect linear function of student-specific incentive strength in 2003.

While separate identification of γ^e and θ relies on the non-monotonic form of the 2003 effort function, note that the precise functional form is not assumed but rather *estimated* using our difference-in-differences strategy that exploits the introduction of NCLB in 2003 as an exogenous shock to incentives.

Identification Argument *with* ABCs Incentives

The key identifying assumption when adding the ABCs component is that ABCs incentives operate *across* schools while NCLB incentives operate *within* schools, thus providing separate identification of the incentive effects of the two schemes.

Here, we rely on an important difference between the designs of NCLB and the ABCs to support this assumption. The ABCs sets only an average school-level growth target, implying that average school-level effort is critical in forming the likelihood of each school passing or failing the scheme. In contrast, NCLB sets a secondary, student-level target (the test score required for subject matter proficiency) in addition to its primary school-level target (the proficiency rate). The student-level target ensures that the *distribution* of student-level effort within a school is relevant in determining the likelihood of school-level success under NCLB.

The argument for separately identifying γ^e from θ is the same as above, while the argument for separate identification of ρ from both γ^e and θ relies on there being significant within-school variation of NCLB incentives. Given the fixed test score proficiency target of NCLB, a direct consequence of large within-school variation of student test scores is that NCLB incentives also vary widely across students within a given school. This ensures that there are a sufficient number of marginal and non-marginal students within all schools with differing values of $\bar{e}^{nclb}_{s2003}$, thus allowing us to separately identify ρ from γ^e and θ.

VI.C. Results for Effort Persistence

The discussion of identification in place, we apply the estimation routine to the sample of fourth grade students in 2004 who have non-missing math scores, third grade effort values from 2003, and counterfactual predicted scores; the resulting sample size consists of 86,237 students.

Table 5 presents the results. The first column provides an estimate of the persistence of effort without accounting for contemporaneous effort incentives. In this case, 50 percent of

the initial effort effect persists one year into the future. As expected, this estimate overstates the true persistence rate, as shown by the estimates in column (2). That is, once we account for contemporaneous NCLB effort, the estimate of γ^e falls to 0.13, implying that only 13 percent of the initial effort effect persists to affect 2004 test scores. The estimate of $\theta = 0.51$ in column (2) implies that the effort response is scaled down by 50 percent in 2004 relative to 2003. The estimate of $\hat{\theta} < 1$ means that the difference between the effort received by marginal and non-marginal students at the average school becomes smaller in 2004 than in 2003, suggesting a smaller redistribution of effort.

Table 5: Maximum Likelihood Parameter Estimates

	(1) Without Contemporaneous NCLB and ABCs Incentives	(2) With Contemporaneous NCLB but Without ABCs Incentives	(3) With Contemporaneous NCLB and ABCs Incentives
γ^e	0.50***	0.13***	0.13***
	(0.02)	(0.03)	(0.02)
θ	-	0.51***	0.47***
	-	(0.02)	(0.02)
ρ	-	-	0.33***
	-	-	(0.06)
μ	0.97***	0.87***	0.30**
	(0.05)	(0.05)	(0.12)
σ^2	20.30***	20.14***	20.13***
	(0.14)	(0.14)	(0.14)
Observations	86,237	86,237	86,237

Notes: This table presents maximum likelihood estimates of variants of equation (26). The sample includes fourth grade students in 2004. The dependent variable in each column in the difference between the realized and counterfactual predicted math score. Standard errors calculated using the Outer-Product of Gradients method appear in parentheses. *** denotes significance at the 1% level;** denotes significance at the 5% level.

Accounting for ABCs incentives (in column (3)) results in an estimate of $\hat{\rho} = 0.33$, which implies that a one standard deviation increase in school-level NCLB effort from 2003 produces a one percent of a standard deviation increase in student-level test scores. Although this is a relatively small effect, the positive and significant estimate of ρ implies that NCLB effort responses from 2003 *strengthened* ABCs incentives for the average school in 2004, leading to student performance gains. Prior to NCLB, the average school passed the ABCs relatively easily, implying that most schools were non-marginal with respect to the ABCs and easily satisfied their targets. The NCLB responses in 2003 made it more difficult for schools to pass the ABCs in 2004, thus making the average school more marginal with respect to the

ABCs and thereby strengthening incentives.[39]

It is important to note that the estimates of γ^e and θ are nearly identical to those obtained from the maximum likelihood routine that does not account for ABCs incentives (the estimate of γ^e is also more precise). This lends credence to the assumption underpinning the identification strategy: NCLB incentives vary within schools while ABCs incentives vary across schools.

VII. Policy Analysis

Building on the model and estimates presented above, this section sets out an approach for placing incentive-based education reforms on a common footing with popular reform proposals in the literature. Broadly speaking, the approach involves computing the benefits in terms of student achievement for different policies, then calculating the respective dollar costs to see which policy is more cost-effective.

Taking the benefits first, we use the model to set the incentives under the incentive-based policy at a level equalizing the implied student performance effects across comparison policies. Then, given our cost-effectiveness focus, we use the estimates of the contemporaneous and persistent effects of effort to derive an upper-bound that, if satisfied by the per-teacher cost of the incentive scheme, would make the incentive scheme more cost-effective than comparison reforms, whose cost has been established in prior research; this step does not require any additional analysis.

Computing the costs of incentive-based policies requires further steps. We require a measure of the cost required to obtain a given amount of extra teacher effort. Given that NCLB is a sanctions-based system that does not offer financial rewards or penalties, we devise a procedure (described below) for estimating the per-teacher costs of extra effort – this combines our model, the connection between NCLB and ABCs incentives, and the ABCs' bonus-payment design. Applying this procedure allows us to back-out the implied value of the monetary equivalent of the NCLB sanction, given the effort response we observe. This, in turn, yields the costs that we seek.

[39]This in the case for the *average* school. Some schools likely found it difficult to pass the ABCs prior to NCLB, implying the NCLB response in 2003 made it even harder to pass in 2004. Such schools would become less marginal and reduce ABCs effort as a result.

This type of cost-effectiveness comparison involving incentive-based reforms has not been carried out in prior work. To illustrate the approach, we compare incentive-based policies for improving teacher effectiveness with policies that seek to improve average teacher productivity by dismissing the lowest-performing teachers. Such 'ability-based' policies have received considerable attention in the recent literature, with Hanushek (2009, 2011), Chetty *et al.* (2014b), and Rothstein (2015) all analyzing policies that dismiss teachers whose value-added falls in the bottom part of the measured distribution (for example, the bottom five percent).[40]

VII.A. Effects on Student Achievement

We begin by placing the two policies on a common footing in terms of their effects on student achievement. Considering ability-based reforms first, Chetty *et al.* (2014b) provide benchmark estimates that indicate that replacing the lowest rated teachers with draws of new teachers results in an average two standard-deviation improvement in teacher ability for that subset. Although our sample differs somewhat, these estimates provide values that can be used in our setting as estimates for the benefits of a similar policy based on teacher replacement.

Our model and estimates offer a way of raising teacher productivity by an equivalent two standard deviations, on average, through an incentive-based reform. The reform we propose has two aspects: setting tougher test score targets for students who are currently non-marginal, and altering the value of the NCLB sanction (or reward under a comparable pecuniary scheme). While NCLB does not offer financial rewards or penalties, we assume teachers assign a monetary equivalent value to the NCLB sanction, which we denoted b^{nclb} in the model and discuss in more detail below.

The required two standard-deviation performance improvement can be attained with an incentive-based reform by first increasing the fraction of marginal students in each classroom from the current sample average, 26 percent, to 86 percent.[41] This is accomplished by making

[40] The literature has also focused on reducing the attrition of the highest rated teachers. However, existing research (Chetty *et al.*, 2014b) suggests that such a focus on the top is a less cost-effective (ability-oriented) reform than replacing the lowest rated teachers.

[41] Non-marginal students with predicted performance above the NCLB accountability target constitute 60 percent of our sample. It is therefore possible to increase the proportion of marginal students in each class by 60 percentage points (from the 26 percent average) by setting more difficult test score targets for non-marginal students.

test score targets $y_g^{T,nclb}$ student-specific (instead of fixed within grades, as under NCLB) in order to make more students marginal with respect to their targets, thus attaching stronger effort incentives to a greater fraction of students in the classroom. To see this, continue to assume that the cost of teacher effort is given by $c(e) = \frac{d}{2}e^2$ (Assumption 3 from the model) and use the equation for optimal effort (equation (11) from the model) to write the effort decision of any teacher j as

$$e_{j2003} = \frac{b^{abcs} f(\Omega_s)}{dN_{gst}} + \frac{\Psi'\big(R_{st}\big)}{dN_{st}} \sum_{i=1}^{N_{j2003}} \frac{h(y_g^{T,nclb} - \gamma y_{ij'g-1s't-1} - a_j - e_{igst})}{N_{j2003}}. \tag{27}$$

When the target $y_g^{T,nclb}$ is set closer to a given student's predicted score, it increases the value of directing an additional unit of effort to that student, captured by $h(y_g^{T,nclb} - \gamma y_{ij'g-1s't-1} - a_j - e_{igst})$. A regime with student-specific targets thus increases each teacher's effort by making a greater fraction of her students marginal. In addition, holding all else constant, making more students marginal with student-specific targets reduces the school's expected pass rate R_{st}, which creates stronger incentives to exert additional effort to all students attending the school. This follows from the marginal benefit of effort being a decreasing function of the school's expected pass rate (that is, $\Psi'\big(R_{st}\big)$ is downward sloping).

In equation (27), both the classroom- and school-level mechanisms work to increase effort, and the reduced-form estimates in Table 3 imply that a 60 percentage-point increase in the proportion of marginal students results in a performance improvement of 0.17 standard deviations of the test score, on average – equivalent to the impact of nearly a one standard-deviation increase in teacher ability.[42]

We achieve the desired additional one standard-deviation performance improvement by raising the value of the NCLB sanction. As mentioned, NCLB does not offer financial rewards or penalties, but we assume teachers assign a monetary equivalent value to the NCLB sanction, which we denote b^{nclb}. In particular, suppose the marginal benefit of improved school-level performance under NLCB is given by $\Psi'\big(R_{st}\big) = b^{nclb}\Lambda(R_{st})$, where b^{nclb} is the monetary equivalent of the benefit derived from better performance under NCLB and $\Lambda(\cdot)$

[42]To see this, use the estimated effort effects and sample sizes for each grade in Panel (a) of Table 3 to construct the following average effect across all grades: $\frac{2144}{7312}(1.55)(0.6) + \frac{2598}{7312}(4.39)(0.6) + \frac{2570}{7312}(2.41)(0.6) = 1.72$ developmental scale points. From Table 1, the standard deviation of the test score across third, fourth, and fifth grade is approximately 10 developmental scale points, implying that the proposed reform results in a $(\frac{1.72}{10})$ 0.17 standard-deviation improvement.

is a decreasing (but always positive) function of the schoool's expected pass rate. Equation (27) then implies that teacher effort is increasing in the monetary equivalent of the sanction, b^{nclb}, making it is possible to raise effort further by increasing this value. Our reduced-form estimates imply that a sanction set at 140 percent of its current value results in an effort-driven performance improvement equivalent to the required extra one standard-deviation improvement in teacher ability.[43]

In this way, the proposed incentive-based policy results in a benefit comparable to replacing the bottom five percent of teachers (in terms of value-added performance) with average draws from the distribution of teachers. The first part of the proposed policy – setting student-specific targets to increase the fraction of marginal students across classrooms – is costless (aside from any administrative disruption). The second part involves setting a different monetary value for the NCLB sanction and is therefore associated with new costs. We quantify the costs in the following subsection.

VII.B. The Cost of Each Reform

Comparing the costs of the ability-based and incentive-based reforms, the ability-based reform creates increased employment risk for teachers throughout the distribution, as estimation error in value-added measures implies that *any* teacher can score in the bottom of the value-added distribution with non-zero probability. Rothstein (2015) finds that compensating teachers for the increased risk requires a mean salary increase across all teachers of 1.4 percent, which amounts to an average increase of $700 per-teacher in North Carolina, where the mean salary is approximately $50,000. We therefore assume that implementing the ability-based reform in our setting comes at an additional cost of $700 per-teacher.

VII.B.1 A Cost-Effectiveness Upper-Bound

Since there is no monetary bonus or sanction under NCLB, the analogous per-teacher cost of raising effort through NCLB-type incentives is not immediately apparent. That said, our estimates of the contemporaneous and persistent effects of teacher effort can be use to derive

[43]Note that a 60 percentage-point increase in the fraction of marginal students achieves a one standard-deviation ability increase in effort (from footnote 42). An average proportion of marginal students of 86 percent thus achieves a 1.43 standard-deviation-of-ability effect (that is, $(0.86/60) * 1 = 1.43$), implying that a two standard-deviation performance gain is achieved by a sanction that is 140 percent (2/1.43) of the current value.

an upper bound that the monetary equivalent of the NCLB sanction would need to be no greater than in order for the proposed incentive reform to be cost-effective.

To calculate that upper bound, we note above that our proposed reform involves setting a value for the NCLB sanction that is 140 percent of the current (unknown) monetary equivalent value. In that case, the resulting effort-driven improvement in teacher performance is equivalent to a two-standard deviation improvement in teacher ability, as we showed above. It is important to note, however, that the improvement under the incentive-based reform is realized throughout the entire distribution of teachers while the improvement under the ability-based reform is realized only within the subset of the bottom five percent of teachers (in terms of value-added). Thus, achieving a comparable average improvement with the incentive-based reform requires a sanction that is only 7 percent of the current value.[44]

If the monetary equivalent of the NCLB sanction is therefore less than $10,000 (i.e. $700/0.07), a sanction that is 7 percent of the current value results in less than a $700 increase in per-teacher costs. This upper bound for the monetary equivalent of the sanction is calculated by setting the contemporaneous effect of the incentive-based reform equal to the contemporaneous effect of the ability-based reform. Our estimates of the persistence rates of teacher effort and ability indicate that the effects of effort decay faster than the effects of ability, implying that the benefits of the incentive-based reform fade out faster in the longer run. We therefore adjust the upper bound of the cost-effectiveness threshold downward to reflect the differential persistence of effort and ability, deriving a long-run upper bound of $3,200 for the monetary equivalent of the NCLB sanction.[45] The incentive-based reform is thus more cost-effective than the ability-based reform in the long run if the NCLB sanction is valued at less than $3,200.

[44]Since the ability-based reform affects only 5 percent of teachers, we divide 140 percent by 20 to arrive at the 7 percent value.

[45]The long-run comparison depends on how effort effects persist beyond one year – an issue we intend to investigate in future work. For now, we note that Figure 7 shows that the effect of ability on test scores four periods into the future is 46.3 percent of the ability effect one period into the future (i.e. 0.19/0.41). Assuming a similar pattern for the effects of effort, the effect four periods ahead amounts to 6 percent (i.e. 0.13*0.463) of the initial effort effect. Thus, four periods forward, the incentive reform achieves 32 percent of the effect of the ability reform (i.e. 0.06/0.19). Scaling the short-run threshold by 32 percent (i.e. $10,000*0.32) thus results in a long-run sanction threshold value of $3,200.

VII.B.2 Estimating the Per-Teacher Monetary Equivalent of the NCLB Sanction

Our current estimates are still uninformative as to the actual costs of the proposed incentive reform. Determining the cost of a performance improvement due to greater effort is challenging (as mentioned) because NCLB incentives are not associated with monetary bonuses or penalties.

To address this challenge, we propose an approach that involves three steps: First, we calculate the degree to which school responses to NCLB lowered the probability of passing the ABCs, relative to the counterfactual scenario in which NCLB was not enacted; the differences in these passing probabilities combined with the ABCs bonus payment determine the expected financial loss each school brought upon itself by responding to NCLB. Second, we use the expected financial loss along with our estimate of how schools respond to changes in ABCs targets from Section VI to determine the relationship between financial incentives and educator effort. Third, we use this relationship and the observed NCLB effort response in 2003 to infer the monetary value educators place on the NCLB sanction, thus allowing us to obtain a cost estimate of the proposed incentive-based reform.

Calculating School-Level Reductions in ABCs Passing Probabilities

We calculate school-level growth scores under the ABCs in 2004 for each school in the counterfactual scenario where NCLB was not enacted, following the aggregation rules set out under the ABCs and substituting in values from our model where appropriate. In particular, since the ABCs measure performance using contemporaneous test scores and set targets using prior scores, we require counterfactual test scores for each student from both 2003 and 2004. Here, we use the predicted score in 2003 as the 2003 test score that would have occurred without NCLB and the counterfactual predicted score from 2004 as the test score that would have occurred in 2004. The differences between 2004 performance and the ABCs target (which uses lagged performance from 2003) for all students are then used to form school-level ABCs growth scores by following the aggregation rules under the program.[46] Using the model notation, the school-level growth score is given by

[46] While the model in Section IV abstracts from some of the detailed rules for calculating school-level ABCs scores, this calculation follows those rules precisely. In particular, we sum the weighted and standardized grade-and-subject-specific average differences between realized and target growth within each school.

$$\sum_{g=3}^{G_s} \sum_{\{i:\ i \in g_{st}\}} \frac{\tilde{y}_{ijsgt} - \alpha \hat{y}_{ij'g-1s't-1}}{N_{gst}}. \tag{28}$$

We also calculate school-level growth scores under the ABCs in 2004 in a scenario in which schools only respond with additional effort in 2003. We do not incorporate 2004 NCLB and ABCs effort responses into the calculation, as we are interested in isolating the reduction in ABCs passing probabilities caused by the initial response to NCLB. In this case, we take the prior score for each student to be the realized prior score and the test score that would have occurred in 2004 to be the sum of the counterfactual predicted score and the persistence of the effort response from 2003. We again calculate the difference between 2004 performance and the ABCs target for each student and use these differences to form school-level ABCs growth scores.[47] In terms of the notation used in the model above, the growth score in this case is given by

$$\sum_{g=3}^{G_s} \sum_{\{i:\ i \in g_{st}\}} \frac{\tilde{y}_{ijgst} + \gamma^e e^{nclb}_{ij'g-1s't-1} - \alpha y_{ij'g-1s't-1}}{N_{gst}}. \tag{29}$$

With school-specific ABCs growth scores, we are able to calculate the school-level probability of passing the ABCs, which is governed in our model by $F(\cdot)$, the cumulative density function of average random fluctuations in test scores at a given school. We represent $F(\cdot)$ using a normal distribution with mean zero and assess the sensitivity of our analysis to a variety of possibilities for the standard deviation of this distribution. Specifically, we let the standard deviation of school-level randomness vary from 0.1 to 1 developmental scale points in increments of 0.1.[48] In each case, we calculate the school-level probability of passing the ABCs with only the 2003 NCLB response, the probability of passing the ABCs assuming that NCLB was never enacted, and the difference between the two, reflecting the degree to which each school lowered the likelihood of passing because of their effort response to NCLB.

[47] For both the scenario in which NCLB never occurred and the one in which we examine NCLB's impact on ABCs passing probabilities, we use realized reading scores as 2004 predicted reading outcomes in the calculations. ABCs reading targets depend on both prior math and reading scores, however, so despite using realized reading scores as 2004 outcomes in both scenarios, we do change the ABCs reading targets to incorporate prior counterfactual math scores where appropriate.

[48] For comparison, the standard deviation of the school-level ABCs score under the counterfactual scenario in which NCLB was not enacted is 0.34 developmental scale points. Although not reported, using even smaller values that are between 0.01 and 0.1 for the standard deviation of randomness does not alter any of our conclusions.

The reduction in the ABCs passing probability due to the NCLB effort response is given by

$$\Delta_{s2004} = F\left(\sum_{g=3}^{G_s}\sum_{\{i:\ i\in g_s\}}\frac{\tilde{y}_{ijsg2004} - \alpha\hat{y}_{ij'g-1s'2003}}{N_{gs2004}}\right) - F\left(\sum_{g=3}^{G_s}\sum_{\{i:\ i\in g_s\}}\frac{\tilde{y}_{ijgs2004} + \gamma^e e^{nclb}_{ij'g-1s'2003} - \alpha y_{ij'g-1s'2003}}{N_{gs2004}}\right). \quad (30)$$

All else equal, the NCLB response decreases the likelihood of ABCs target attainment because the persistence rate of effort (γ^e) is lower than the persistence rate of non-effort inputs. Since the ABCs program does not distinguish between sources of test score gains, ABCs targets increase at higher rate than the stock of knowledge with which students enter the 2004 academic year, making the targets more difficult to attain than in past years. The deviation between the two persistence rates is critical to our method for inferring the monetary equivalent value of the NCLB sanction, as it ensures that NCLB responses resulted in expected financial losses (and associated effort responses) under the ABCs.

Panel (a) of Table 6 provides a summary of the school-level passing probabilities under each counterfactual scenario. For a 0.1 scale-point standard deviation of noise, the average difference between the two passing probabilities is 19 percentage points, while it is 8 percentage points for a 1 scale-point standard deviation of noise, and monotonically decreasing in between.

The Relationship between School Effort and Expected Financial Loss

With the reduction in ABCs passing probabilities for each school in hand, we quantify the relationship between expected financial losses and school-level effort in two steps. Central to our approach is the implication from our model that average school-level NCLB effort from 2003 is the primary determinant of the NCLB-induced change to each school's ABCs passing probability and, correspondingly, each school's expected financial loss in 2004.

As our first step, we note that the analysis in Section VI already establishes an indirect relationship between expected financial losses under the ABCs and teacher effort. In particular, we showed that student test scores in 2004 depend positively ($\hat{\rho} = 0.33$) on average school-level NCLB effort from 2003. We interpret this relationship as reflecting an *indirect* relationship between expected financial losses under the ABCs and the associated effort

Table 6: School-Level ABCs Passing Probabilities and Estimates of NCLB Effects

Standard Deviation of School-Level Error Term	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Panel (a): Counterfactual School-Level Passing Probabilities										
Average Passing Probability if NCLB was only in effect in 2003	0.73	0.71	0.68	0.65	0.64	0.62	0.61	0.60	0.59	0.58
Average Passing Probability if NCLB was not enacted	0.92	0.89	0.85	0.80	0.77	0.74	0.71	0.69	0.67	0.66
Average Difference in Passing Probabilities	-0.19	-0.18	-0.17	-0.15	-0.13	-0.12	-0.10	-0.09	-0.08	-0.08
Panel (b): Estimates from Regression of Difference in 2004 Passing Probabilities on 2003 School-Level NCLB Effort										
Effect of School-level NCLB Effort	-0.29 (0.03)	-0.24 (0.03)	-0.20 (0.02)	-0.17 (0.02)	-0.14 (0.01)	-0.12 (0.01)	-0.11 (0.01)	-0.10 (0.01)	-0.09 (0.01)	-0.08 (0.01)
Observations	1,250	1,250	1,250	1,250	1,250	1,250	1,250	1,250	1,250	1,250

Notes: The unit of observation is a school in 2004. In panel (a), we calculate the probability of passing the ABCs for each school, assuming NCLB only operated in 2003 and assuming NCLB was never enacted. The average passing probability for each scenario is reported in rows (1) and (2), respectively, under each possible value for the standard deviation of the school-level error term, as listed in the column headings. In row (3), we calculate the difference in passing probabilities for each school across the two scenarios and report the average of these differences. In panel (b), we regress the difference in passing probabilities on average school-level NCLB effort from 2003, and report the resulting estimates under each value of the school-level error term. Standard errors are reported in parenthesis. Each coefficient is significant at the 1 percent level.

responses, as average school-level effort from 2003 determines the change to ABCs passing probabilities but, conditional on student-specific effort from 2003, should not have any direct effect on student test scores in 2004. As our second step, we estimate the relationship between school-level effort from 2003 and expected financial losses, using the resulting estimate to scale the indirect effect of school-level effort on student test scores ($\hat{\rho} = 0.33$). Scaling the effect of lagged school-level effort on tests scores by the effect of lagged school-level effort on expected financial losses results in the *direct* effect of expected financial losses on test scores, which we assume is mediated by teacher effort. The intuition for the identification strategy is presented in a diagram in Figure 9.

The second step – estimating the effect of school-level NCLB effort on expected financial losses – is carried out by regressing the change in the school-level likelihood of passing the ABCs in 2004 on average school-level effort in the prior year:

$$\Delta_{s2004} = \alpha + \beta \bar{e}^{nclb}_{s2003} + \epsilon_{s2004}. \tag{31}$$

The estimate $\hat{\beta}$ governs the magnitude by which a one-unit increase in school-level effort in 2003 lowers the likelihood of ABCs target attainment in 2004. Panel (b) of Table 6 reports the estimated coefficients from equation (31). The coefficient ranges from -0.29, when the standard deviation of noise is assumed to be 0.1, to -0.08, when the standard deviation is assumed to be 1. A one unit increase in average school-level effort in 2003 thus reduces the probability of passing the ABCs by a value between 8 and 29 percentage points.

Multiplying these values by the ABCs 'high-growth' bonus payment of \$1500 per teacher and scaling the estimate of $\hat{\rho} = 0.33$ by the result provides an estimate of the direct relationship between expected financial losses and effort-driven test score improvements in 2004. In particular, a \$1 expected financial loss under the ABCs causes an effort-driven increase in average student test scores that is between 0.0007 and 0.0025 developmental scale points.

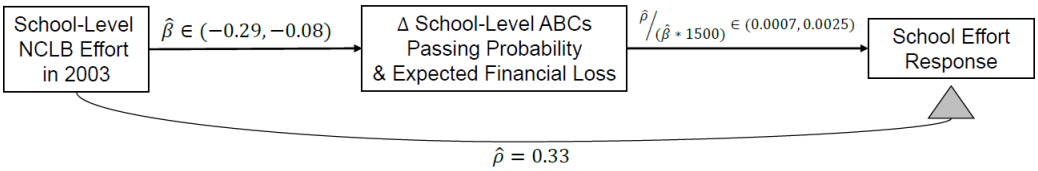


Figure 9: Identifying the Effect of Expected Financial Loss on School Effort

Calculating the Implied Valuation of the NCLB Sanction

We use these estimates to infer the value educators place on the NCLB sanction by relating the observed effort responses to NCLB to the estimated relationship between expected financial losses and effort. The average school-level NCLB effort gain in 2003 is 1.96 developmental scale points. Combining the average effort response under NCLB with the estimates above suggests that the NCLB sanction is valued between \$784 (i.e 1.96/0.0025) and \$2,800 (i.e. 1.96/0.0007) per-teacher. Based on the observed NCLB response as well as the resulting variation in ABCs targets, and the ABCs financial bonus, we therefore estimate an upper-bound value for the NCLB sanction of \$2,800.[49]

[49]We implicitly assume that educators respond the same way to a given expected financial loss irrespective of whether the loss stems from the ABCs or NCLB. Conceptually, our first-order condition for optimal effort in equation (11) shows that this assumption rests on there being one effort choice governed by one cost function. Manipulating either the ABCs or NCLB object on the left-hand-side of equation (11) by a given amount then results in the same change to effort decisions.

VII.C. A Cost-Effectiveness Comparison of the Two of Reforms

These cost calculations suggest that a program similar to NCLB that offers a bonus payment between $784 and $2,800 per teacher when schools succeed could result in similar effort responses to those observed under NCLB. These figures help determine the approximate cost of such an incentive-based scheme, making it directly comparable to other policy reforms discussed in prior work. In the current analysis, we compare our proposed incentive-based reform to a common variant of the ability-based reform, which involves replacing the bottom five percent of teachers.

Our estimates of the contemporaneous and persistent effects of teacher effort imply a target long-run upper bound of $3,200 that the incentive reform would need to satisfy for the monetary equivalent of the NCLB sanction in order for the proposed incentive reform to be cost-effective relative to the ability reform. The (upper bound) estimate of $2,800 that we estimate for the per-teacher cost of NCLB incentives falls below this threshold, suggesting that the proposed incentive-based reform is potentially more cost-effective than the ability-based reform. Recall that we only require a sanction set at 7 percent of its current value for the incentive-based reform to achieve a benefit comparable to that under the ability-based reform in the short run. Taking $2,800 as the true value of the NCLB sanction, 7 percent of the sanction amounts $196 per teacher. Scaling this value to account for the effects of effort amounting to only 32 percent of the long-run effects of ability implies that the required per-teacher cost of the incentive reform amounts to $612.50, or 87.5 percent of the $700 cost required to compensate teachers for increased employment risk under the dismissal-based reform.

As a final comparison between the two policies, recall that the ability-based reform affects only a pre-determined fraction of teachers who fall in the bottom of the teacher performance distribution. Our calculations of the costs and benefits of the incentive-based reform are scaled to account for this feature of the ability-based reform, thus allowing us to compare the two policies directly. However, it is important to note that the incentive-based reform can be scaled to affect the effort decisions of the full distribution of teachers, while the ability-based reform is limited in this regard.[50] Thus, the incentive-based reform potentially

[50]To see why, note that the ability-based reform entails firing the bottom 5 percent of teachers and replacing them with a random draw from the full distribution. In expectation, a newly drawn teacher's value-added is equal to

provides a viable, cost-effective policy lever for improving average teacher performance.

VIII. Conclusion

Incentive-based education policies have become increasingly widespread over the past two decades. Yet how they compare with alternative types of education policy has remained under-explored, in part because of a lack of a framework for quantitative comparison. This paper proposes such a framework, allowing the cost-effectiveness of incentive-based policies to be computed alongside that of rival policies for the first time.

On the benefit side of the cost-effectiveness calculation, we presented an approach permitting us to separate out teacher effort, which is responsive to education incentives, from teacher ability, which is not. This allows us to gauge the respective impacts of effort and ability on contemporary scores. Further, we measured the extent to which of these two potentially important education inputs persist differentially.

Central to our approach was a novel identification strategy taking advantage of a natural experiment associated with the introduction of a federal accountability program in a setting – the state of North Carolina – where accountability incentives already operated. Specifically, we drew on the proficiency-count design of NCLB to construct a measure of incentive strength for each teacher, showing a positive relationship between teacher value-added and this measure in the year NCLB was introduced but not in prior years. We exploited these differential relationships over time to separate teacher quality into teacher ability and the effort response associated with NCLB. Then to evaluate the persistence of teacher ability and effort, we drew on a simple model of school-decision making to develop a structural estimation strategy, which allowed us to identify the persistence of effort separately from the effects of contemporaneous incentives. Here, we found that effort has a significant positive effect on future test scores, but that the effect of effort persists at approximately 32 percent of the ability effect.

Together, our findings indicate that the effort margin is first-order: teacher effort is both

the mean value-added, implying that the marginal benefit of such a policy is declining as the dismissed teachers come from progressively higher positions in the value-added distribution. One would eventually face the prospect of dismissing teachers with above average value-added. At this point, the policy yields negative returns and forces a switch to policies geared toward retaining high-performing teachers, which the literature has already shown are inferior to policies based on dismissal.

a productive input and one that is responsive to incentive variation in a systematic way, with longer-term benefits for students.

Based on these estimates and the model, we were able to explore the cost-effectiveness of incentive reforms alongside alternative education reforms discussed in the literature. Our illustrative analysis indicated that using formal incentives constitutes a viable alternative means of accomplishing the goal of raising student and school performance, and can be more cost-effective than competing ability-based reforms.

The general approach serves to open up a fuller comparison of the cost-effectiveness of alternative policies, based on further refinements to the estimation approach we develop – for instance, looking at the longer-run persistence of effort. These are areas we are exploring in ongoing research.

References

Ahn, Thomas and Jacob Vigdor. 2014. "The Impact of No Child Left Behind's Accountability Sanctions on School Performance: Regression Discontinuity Evidence from North Carolina." National Bureau of Economic Research Working Paper 20657.

Bacher-Hicks, Andrew, Thomas J. Kane, and Douglas O. Staiger. 2014. "Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles." National Bureau of Economic Research Working Paper 20657.

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

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2011. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." National Bureau of Economic Research Working Paper 17699.

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

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review*, 104(9): 2633-2679.

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

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.

Goldhaber, Dan and Michael Hansen. 2013. "Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance." *Economica*, 80: 589-612.

Hanushek, Eric A. 2009. "Teacher Deselection." in *Creating a New Teaching Profession*, ed. Dan Goldhaber and Jane Hannaway, 16580. Washington, DC: Urban Institute Press.

Hanushek, Eric A. 2011. "The Economic Value of Higher Teacher Quality." *Economics of Education Review*, 30: 466-479.

Hanushek, Eric A. and Margaret E. Raymond. 2005. "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management*. 24(2): 297-327.

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

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

Kane, Thomas J. and Douglas O. Staiger. 2002. "Volatility in School Test Scores: Implications for Test-Based Accountability Systems," in *Brookings Papers on Education Policy*, edited by D. Ravitch. Washington, DC: Brookings Institution Press.

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

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. 'Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." *Report Prepared for the Measuring Effective Teaching Project.*

Kane, Thomas J. and Douglas O. Staiger. 2014. "Making Decisions with Imprecise Performance Measures: The Relationship Between Annual Student Achievement Gains and a Teacher's Career Value-Added." Chapter 5 in Kane, T.J., Kerr, K.A. and Pianta, R.C. Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project. San Francisco.

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.

Lavy, Victor. 2002. "Evaluating the Effect of Teachers Group Performance Incentives on Pupil Achievement." *Journal of Political Economy*. 110(6): 1286-1317.

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

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

Macartney, Hugh, Robert McMillan, and Uros Petronijevic. 2015. "Incentive Design in Education: An Empirical Analysis." National Bureau of Economic Research Working Paper 21835.

McCaffrey, Daniel F., J.R. Lockwood, Kata Mihaly, and Tim R. Sass. 2012. "A Review of Stata Routines for Fixed Effects Estimation in Normal Linear Models." *Stata Journal*, 12(3).

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.

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

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics*, 125(1): 175-214.

Rothstein, Jesse. 2014. "Revisiting the Impacts of Teachers." University of California, Berkeley Working Paper.

Rothstein, Jesse. 2015. "Teacher Quality Policy When Supply Matters." *American Economic Review*, 105(1): 100-130.

Todd, Petra and Kenneth Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal*. 113(February): F3-F33.

Vigdor, Jacob. 2009. Teacher Salary Bonuses in North Carolina. M.G. Springer, ed., Performance Incentives: Their Growing Impact on American K-12 Education. Washington: Brookings Institution Press.

Wiswall, Matthew J. 2013 "The Dynamics of Teacher Quality." *Journal of Public Economics*, 100: 61-78.

A. Mathematical Appendix

Proof of Proposition 1

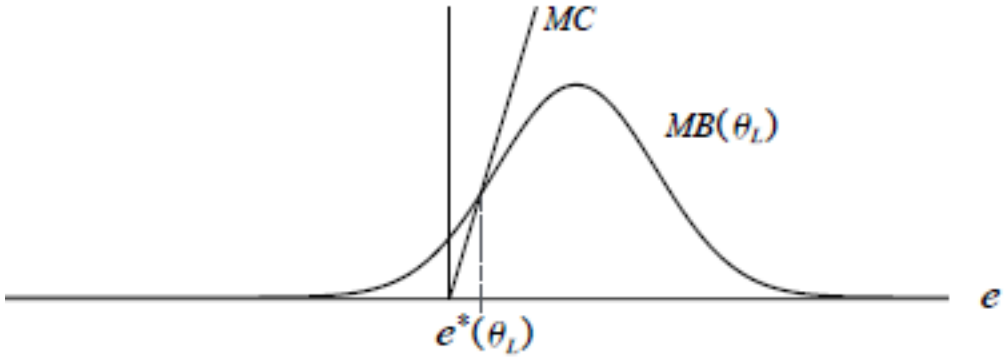
The first-order condition for teacher effort (equation (11)) can be written as

$$b^{abcs} f(\Gamma_s^*) \frac{1}{N_{gst}} + \Psi'\big(R_{st}\big) \frac{h(\pi_i - e_{ijgst}^{nclb})}{N_{st}} = c'(e_{ijgst}),$$

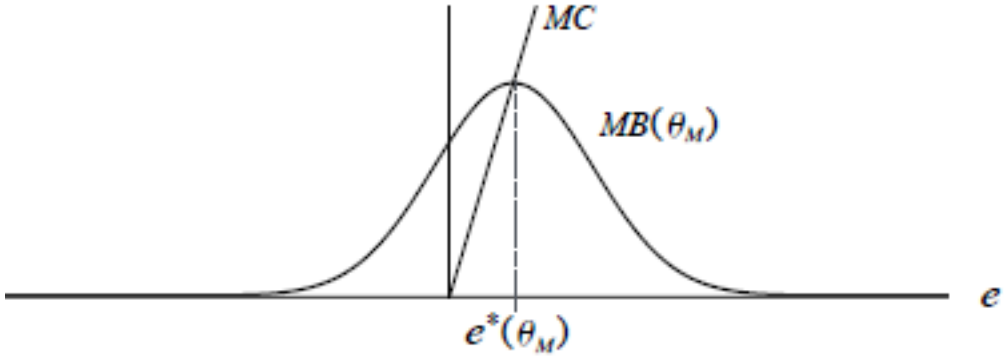
where $\pi_i = y_g^{T,nclb} - \hat{y}_{ijgst}$. The ABCs component of effort $b^{abcs} f(\Gamma_s^*) \frac{1}{N_{gst}}$ and the school-level component of NCLB effort $\Psi'\big(R_{st}\big)$ are common to all students within a school and are therefore ignored throughout the proof.

The proof focuses on the student-level component of NCLB effort, $h(\pi_i - e_{ijgst}^{nclb})$, and proceeds graphically with the aid of Figure 10 below, which maintains the assumptions that $h(\cdot)$ is a unimodal distribution and that the marginal cost of effort is linear. Panel (b) of the figure depicts the effort solution to the first-order condition for the intermediate value of $\hat{y}^*$ that results in maximum NCLB effort. The value $\hat{y}^*$ is such that the solution to the first-order condition also satisfies $e^* = y_g^{T,nclb} - \hat{y}^*$. Here, the distance between the target and the predicted score, $\pi_i^* = y_g^{T,nclb} - \hat{y}^*$, is relatively small in absolute terms. As $\hat{y}$ decreases below $\hat{y}^*$, increasing π_i and moving $h(\cdot)$ to the right, the intersection with the marginal cost curve occurs at a lower level of effort, as shown in panel (a). Panel (c) shows that increasing $\hat{y}$ far above $\hat{y}^*$ decreases π_i (but increases it in absolute terms relative to π^*), shifts $h(\cdot)$ left, and also results in a level of effort lower than that under $\hat{y}^*$.

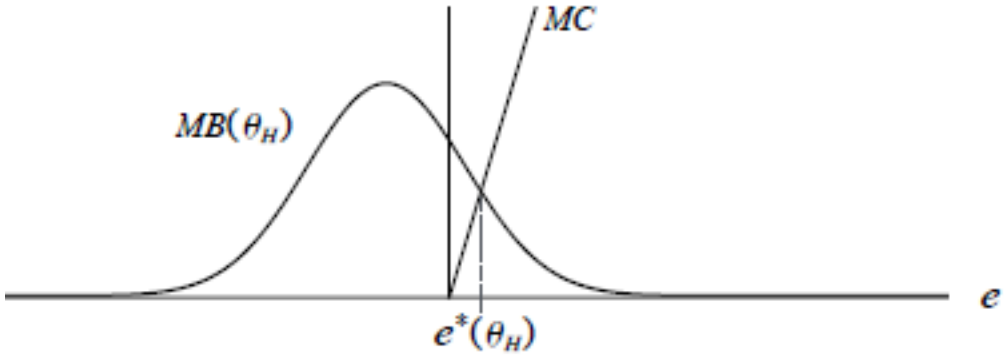
Thus, students with either high or low predicted scores receive less effort than students with intermediate values of predicted scores. ■



(a) Low θ relative to y^T



(b) Intermediate θ relative to y^T



(c) High θ relative to y^T

Figure 10: Graphical Depiction of NCLB Effort