NBER WORKING PAPER SERIES

TEACHER PERFORMANCE AND ACCOUNTABILITY INCENTIVES

Hugh Macartney
Robert McMillan
Uros Petronijevic

Working Paper 24747
http://www.nber.org/papers/w24747

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2018

We would like to thank Raj Chetty, Damon Clark, Peter Cziraki, John Friedman, Caroline Hoxby, Magne Mogstad, Louis-Philippe Morin, Juan Carlos Suárez Serrato, Hammad Shaikh, Brooklynn Zhu, and seminar participants at Arizona State University, Chicago Harris, Columbia University, Duke University, McMaster University, SUNY Buffalo, UC Irvine, the University of Ottawa, Wilfrid Laurier University, the NBER Public Economics Fall 2015 meeting, the NBER Economics of Education Fall 2016 meeting, and the Northwestern Interactions Conference for helpful comments and suggestions. Mike Gilraine provided outstanding research assistance. Financial support from SSHRC and the University of Toronto Mississauga is gratefully acknowledged. All remaining errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Teacher Performance and Accountability Incentives
Hugh Macartney, Robert McMillan, and Uros Petronijevic
NBER Working Paper No. 24747
June 2018
JEL No. I21,J24,M52

ABSTRACT

This paper documents a new empirical regularity: teacher value-added increases within-teacher when accountability incentives are strengthened. That finding motivates a strategy to separate value-added into incentive-varying teacher effort and incentive-invariant teacher ability, combining rich longitudinal data with exogenous incentive-policy variation. Our estimates indicate that teacher effort and ability both raise current and future test scores, with ability having stronger effects. These estimates feed into a framework for comparing the cost-effectiveness of alternative education policies. For illustration, we show incentive-oriented reforms can outperform policies targeting teacher ability, given their potential to influence all teachers rather than a subset.

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

Uros Petronijevic
York University
Department of Economics
Vari Hall
4700 Keele Street
Toronto, ON M3J 1P3
CANADA
upetroni@yorku.ca

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

I Introduction

Education remains central to the public policy debate, not only given its capacity for building fundamental skills, but also because of a pervasive sense that many public schools continue to underperform. The quest 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, the foundational 2014 papers by Chetty, Friedman and Rockoff (henceforth CFR). These types of measure have become the cornerstone of policy interventions that include, somewhat controversially, dismissing low-VA teachers. A second prominent strand studies the impact of accountability incentives on student and school performance – an issue not taken up in the VA literature. In that vein, a number of persuasive papers show that accountability schemes, which have become increasingly widespread in the United States over the past two decades, have succeeded in improving student achievement in a variety of settings (see Figlio and Loeb (2011) for a comprehensive survey).

This paper brings these two literatures together by studying the impact of accountability incentives on measured teacher performance. It makes three related contributions. First, we demonstrate that teacher VA *increases* in incentive strength, using rich administrative data covering all North Carolina public school students over time. Our incentive measure exploits exogenous variation arising from the introduction of the federal "No Child Left Behind" (NCLB) accountability system. As is well-appreciated in the literature (see Reback 2008, for instance), proficiency schemes such as NCLB make students matter differentially depending on how marginal they are (captured by the closeness of their predicted test score to a fixed performance target). Drawing on this insight, we compare teachers who teach classrooms with higher versus lower proportions of marginal students, both before and after the incentive reform was implemented in North Carolina in the 2002-03 school year. Our finding that teacher VA increases in incentives is robust to considering across- and within-teacher variation as well as alternative incentive strength definitions. Further, the evidence suggests a teacher *effort* response, as the reform did not cause school principals to engage in potentially costly within-school teacher reassignments, nor to alter class sizes.

Building on that suggestive evidence, we make a formal distinction between two inputs into education production: teacher effects that are responsive to the incentive environment, which we label 'teacher effort,' and those that are invariant to it, labelled 'teacher ability.' In what follows, we will take effort to be any incentive-related action that raises scores – for instance, exerting more effort in the classroom, devoting more time to lesson planning, or organizing extra tutoring sessions after school. Our measure of ability will equal the component of teacher VA that does not change over time, conditioning on teacher experience, even though 'ability' in common parlance is sometimes taken to be amenable to change (through teacher training, for example).

As our second contribution, we draw on the ability/effort distinction and set out an estimation approach for identifying contemporaneous teacher ability and effort in terms of test scores, as well as the persistent effects of each. The approach invokes minimal assumptions: linearity of the production technology (which serves as a reasonable first-order approximation to the true technology), education inputs being cumulative in their impact over time, and educators responding to NCLB incentives in a similar way across years by directing relatively more effort to students predicted to score near the proficiency threshold.

Implementing the approach, we estimate the contemporaneous ability of teachers who are observed before and after the reform using standard VA methods and pre-reform data.[1] We then identify contemporaneous effort from the correlation of post-reform VA with incentive strength (captured by the proportion of marginal students in the classroom), conditioning on the estimated pre-reform ability measure and experience. The estimates indicate that a one standard deviation increase in ability raises scores by 0.18 SD, compared to 0.05 SD for a one standard deviation increase in effort.

Having recovered these contemporaneous teacher ability and effort measures, we investigate the extent to which each input persists in determining a student's test scores in future. To estimate the persistence of teacher ability, we use data from the pre-reform period and adapt a well-established method from the prior literature for determining the persistence of teacher effects, regressing future test scores on our measure of contemporaneous teacher ability. In line with prior work (CFR 2014b, for instance), we find that approximately 40 percent of the initial

[1] VA estimates 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; CFR 2014a; and Kane and Staiger 2014.)

effect of teacher ability persists after one year and that 20 percent remains after four years.

Identifying effort persistence is more challenging. The ideal experiment would involve a single one-time incentive reform that resulted in an immediate effort response, with no correlated responses in the future, allowing us to estimate the persistence of effort using a strategy similar to the approach we follow for ability. In practice, several issues arise. First, incentives to exert effort under NCLB are strongly correlated over time, so to avoid overstating the persistent effects of lagged effort, it is important to account for the impact of contemporaneous effort on current scores. Second, contemporaneous effort itself depends on the persistence parameter we wish to estimate, given that educators make effort decisions based on expected student performance and will use any persistence in effort to form these predictions. Third, we must allow for induced changes to a concurrent state-level accountability program – the ABCs of Public Education – due to the introduction of NCLB (described below) to avoid confounding school-level ABCs-related improvements with student-level effort persistence. We account for these factors directly in our estimation approach, built around a transparent education production technology.

The estimates of the persistence of effort indicate that 10 percent of the initial effort effect persists one year ahead, which amounts to approximately 25 percent of the one-year persistence of teacher ability. 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. Further, we show that not accounting for the test score effects of contemporaneous effort decisions would result in a significant overestimate in the persistence of effort.

Our estimates indicate that teacher effort is both a productive input and one that responds systematically to incentive variation, with longer-term benefits for students. This prompts the natural question: To what extent can education policy makers use incentives to harness teacher effort in a productive way?

We explore that issue as part of our third contribution – using these estimates and the technology to develop a framework for comparing the cost effectiveness of alternative education policies, including those that alter accountability incentives. The estimated effects of teacher ability and effort on student achievement in the short and longer term are relevant for computing policy benefits. On the cost side, given that NCLB incentives are sanctions-based, we use our framework to monetize the value of those sanctions.[2] Doing so allows the costs of incentive

[2]As described below, we take advantage of the exogenous loss in monetary rewards under the ABCs when schools responded to the introduction of NCLB.

reforms to be placed alongside various alternatives.

We present an illustrative policy comparison using the cost-effectiveness framework. The prior literature (notably Hanushek 2009, 2011; and CFR 2014b) has considered altering the teacher ability distribution as a policy lever to raise student scores, specifically replacing teachers whose value-added falls in the bottom five percent of the measured distribution. Given our core finding that incentives matter when measuring the effects of teachers, changing formal incentives offers a viable alternative for raising student and school performance. While our estimates indicate that variation in teacher ability has stronger impacts on outcomes than variation in effort (both in the present and persisting into the future), effort-based effects are substantially less costly to achieve, given that teacher effort can be altered right across the teacher VA distribution. As a result, we show that incentive-oriented reforms can be competitive with policies targeting teacher ability, coming out ahead in a variety of plausible cases. More generally, our framework allows us to extend the policy discussion by providing researchers and policy makers a means to compare the cost effectiveness of incentive-based reforms with various other popular education policies for the first time.

The remainder of the paper is organized as follows: The next section describes the accountability programs we use for identification, as well as the North Carolina administrative data. Section III presents descriptive evidence that motivates the main approach; Section IV sets out the education production technology that forms the basis for our empirical analysis; and Section V outlines our empirical strategy for decomposing ability and effort contemporaneously, along with the results from that exercise. In Section VI, we describe how we estimate the persistent effects of effort (versus ability) and the associated findings, followed by the cost-effectiveness framework in Section VII, which we use to conduct relevant policy comparisons. Section VIII concludes.

II Institutional Background and Data

Our analysis focuses on North Carolina, a state that offers wide variation in performance incentives across teachers and schools, as well as rich administrative data covering all public schools, their teachers and students, followed over time.

II.A Accountability Incentives

Incentive variation in the state arises from two separate accountability regimes. Our main focus is on NCLB, which was implemented in North Carolina for the 2002-03 school year following the passage of the federal No Child Left Behind Act in 2001. NCLB introduced student-level test score targets that students needed to achieve in order to be deemed proficient. As is well appreciated, such thresholds create incentives for teachers to direct relatively more effort toward students likely to score close to the target. NCLB also introduced a 'primary' target, requiring that a fixed percentage of students in the school be proficient on state tests. In a North Carolina setting, this school-level target was set low in the performance distribution and did not generate marked differences in behavior across schools, so we focus on the student-level targets.[3] When schools failed to satisfy NCLB requirements, they were subject to sanctions that became more severe over time in the event of repeated failure.

The second of the accountability regimes – the state's ABCs of Public Education – was implemented in the 1996-97 school year for all schools serving students in kindergarten through eighth grade. Under the ABCs, each school was assigned an average growth target, depending on prior student performance and a constant level of expected growth. If average student test scores at the school exceeded the target, the ABCs paid a monetary bonus to all teachers and the principal.[4] In contrast to NCLB, only a *school-level* performance target was set under the ABCs, requiring that each school achieve sufficiently high overall growth, irrespective of where that growth was concentrated in the underlying student-level achievement distribution. Teachers were therefore not incentivized to direct effort across individual students in a differential way under the ABCs – a feature we take advantage of below.

II.B Data and Descriptive Statistics

Our analysis uses 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 all third through eighth grade public school students, encrypted identifiers for students and teachers, and unencrypted school identifiers. Thus, students can be tracked

[3]The same given percentage of students in each of nine demographic subgroups was also required to achieve proficiency. We do not exploit that separate source of variation in the current analysis.

[4]We will use the fact that the ABCs paid monetary rewards in the cost-effectiveness analysis below. (Macartney 2016 provides additional detail about the ABCs program.)

over time, and linked to a teacher and school in any given year. We provide an overview of the data here, while Appendix A gives more detail.

The main sample runs from the 1996-97 to the 2004-05 academic year and covers over 2.5 million student-year observations. Table 1 provides summary statistics. In terms of performance measures, we focus on end-of-grade (EOG) test scores for students in third through fifth grade who can be linked reliably to teachers following the linking procedure used in prior work.[5] These scores 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.[6] 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 longitudinal nature of the data set enables us to construct growth score measures for both mathematics and reading, based on within-student gains. Student gain scores are (as noted above) 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. It is apparent from the table that mathematics and reading growth is positive on average in both subjects across grades, with the largest gains in both subjects occurring in the earlier grades.[7]

The data set includes information about individual students' gender, race, disability status, limited English-proficiency classification, free lunch eligibility, and grade progression – demographic characteristics that serve as useful control variables. 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 lunches. Around 25 percent of students have college-educated parents, and very small fractions of students repeat a grade.

[5] We restrict attention to students in these grades, following previous studies that use the NCERDC data, given that the teacher recorded as the test proctor is typically the teacher who taught the students throughout the year.

[6] The appendix describes how we accommodate changes to the developmental scales that the student mathematics and reading test scores are measured using.

[7] The table also reports 'future' mathematics and reading scores – the scores we observe for our sample of third through fifth grade students when they are in sixth, seventh, and eight grades, which are used when measuring the persistent effects of teacher ability and effort below.

III Descriptive Evidence

We now present evidence that motivates the subsequent empirical analysis, in two parts: first, evidence of an incentive response to the introduction of NCLB, consistent with there being a targeted effort increase by teachers, and second, correlations suggesting that teacher value-added is responsive to incentives.

Our strategy for uncovering teacher effort draws on the introduction of NCLB in 2002-03, treating that as an exogenous shock to educators' performance incentives. As noted in the Introduction, proficiency-count systems like NCLB are known to provide educators with strong incentives to direct resources to students who are on the margin of passing relative to the scheme's fixed proficiency target, potentially at the expense of those in the tails of the predicted test score distribution (see, for example, Reback 2008, Neal and Schanzenbach 2010, and Deming *et al.* 2016).[8] Thus, the introduction of NCLB should give rise to an inverted U-shaped increase in test scores, centered roughly around the passing threshold.

That prediction is borne out empirically, as Figure 1 indicates. The horizontal axis measures the gap between the student score predicted on the basis of a rich set of covariates (including past student performance) and the fixed target under NCLB – the gap provides an intuitive measure of incentive strength.[9] The vertical axis plots the gap between actual and predicted scores for each student (averaged across all students placed in suitable incentive strength bins), comparing before and after NCLB was introduced in 2002-03. As expected, students experienced no gain over their predicted scores at any point of the predicted score distribution in the *pre*-NCLB period: the profile is roughly flat. In contrast, following NCLB's introduction, we see a pronounced incentive response, captured by the hump in score gains over predicted scores. These were highest for students predicted to score close to the test-score proficiency threshold, with the gains over predicted scores decreasing as one moves further away from that threshold on either side.

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 show that teacher-year fixed effects, which are commonly used to measure teacher effectiveness, actually

[8] Marginal students provide teachers with the greatest expected return to additional effort. In contrast, students predicted to score far above the target are likely to pass anyway, while those predicted to score far below require a prohibitively costly amount of effort to pass.

[9] We discuss the construction of predicted student scores in Appendix D.II.

covary with a simple proxy for NCLB incentive strength in 2002-03.[10] We define a student as 'marginal' if she is predicted to score within four developmental scale points on either side of the proficiency cutoff, and calculate the fraction of students in each classroom who are marginal in that sense – below (see Section V), we will show the results are robust to various alternative choices of cutoff for defining marginal students. 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 classroom.

As Figure 2 shows, this is what we find: teacher-year fixed effects are positively correlated with the proportion of marginal students within a classroom. The relationship is significant (at the one percent level) and positive in each grade in 2002-03, with a one standard deviation increase in the classroom proportion of marginal students being associated with 7, 17, and 11 percent standard deviation increases in teacher-year VA in third, fourth, and fifth grade, respectively. These raw data patterns suggest that NCLB may have caused teachers to exert more effort due to 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. We document that pattern in Section V.[11]

IV Conceptual Framework

Our estimation approach is underpinned by a conceptual framework centering on the education production technology. From an empirical standpoint, estimating the parameters of this technology presents important challenges, given its underlying specification is unknown and many inputs are unobserved, even in the most comprehensive administrative datasets. Our approach to these empirical challenges (developed in the two following sections) exploits plausible sources of policy variation and rich longitudinal data, in combination with some minimal structure.

We focus on the contemporaneous and persistent effects of two key inputs, teacher ability and teacher effort, studying their separate impacts on measured output given by test scores. Both are typically *unobserved*, so our strategies for identifying each will be central to the analysis.

We make two assumptions at the outset:

[10] Appendix B contains a detailed discussion of how teacher-year fixed effects are estimated.

[11] There, we also argue that more than simple correlational plots are required to account for a confounding negative correlation between marginal student presence and teacher ability that arises from the sorting of students to teachers based on ability. The estimates in that section will account for sorting.

Assumption 1: The education production technology is linear in its inputs, with an additive error.

In this regard, we follow convention in the education literature.

Assumption 2: Inputs have a cumulative effect on output.

This second assumption is especially relevant in the case of education, where education investments serve to increase the stock of knowledge over time. We will treat time discretely, corresponding to our yearly data. The current academic year is denoted by t, and a student's grade is indexed by $g \in \{0, 1, 2, ...\}$, where the first year of formal schooling, kindergarten, is represented by $g = 0$. We can then refer to any academic year, past or present, by defining a 'lag' index τ that takes on integer values from 0 up to g.

Reflecting these two assumptions, we consider the following underlying representation of the technology, making explicit how each of the two inputs affects student learning,[12] both in the current year t and in each prior academic year:[13]

$$y_{ijgst} = \sum_{0 \leq \tau \leq g} \left[\gamma_\tau^a a_{j(i,t-\tau)} + \gamma_\tau^e e_{j(i,t-\tau)} \right] + \nu_{igjst} \,. \tag{1}$$

Equation (1) describes the test score of student i, who is assigned (exogenously) to teacher j in grade g at school s in year t. That score is allowed to be a function of the full history of relevant school inputs, extending back to the first year the student was in school (in period $t - g$). Input $a_{j(i,t-\tau)}$ is the ability of student i's teacher in year $t - \tau$, $e_{j(i,t-\tau)}$ is the effort of the teacher in that year given the prevailing incentives (allowed to be student-specific, as under NCLB), and ν_{igjst} is an additive error term. While teacher ability and teacher effort are measured in the same developmental scale units, the parameterization of the input productivities in (1) allows teacher ability to have a different impact on scores than teacher effort – something we test below. Because the unobserved contemporaneous inputs and parameters cannot be separately identified, we normalize the contemporaneous effects of a one-unit change in teacher ability and a one-unit change in teacher effort to be equal (so $\gamma_0^a = \gamma_0^e = 1$), while still allowing for potential

[12] To help fix ideas, we abstract from all other inputs in this section. We also take teacher and student assignments to classrooms as given – the empirics below will address non-random sorting.

[13] Most studies examine the effects of contemporaneous inputs, accounting for the history of past inputs by controlling for prior test scores. We will show how our formulation can be expressed on that standard basis, and what that implies, later in the section.

differences in the persistent effects of ability and effort (that is, $\gamma_\tau^a \neq \gamma_\tau^e$ for $\tau > 0$).

Bringing equation (1) to the data in an ideal setting, one could identify teacher ability and effort, $a_{j(i,t-\tau)}$ and $e_{j(i,t-\tau)}$, separately for each teacher, both in the current year and in prior years, and also estimate the persistent effects of past ability and effort on test scores, captured by the full set of parameters $\{\gamma_\tau^a, \gamma_\tau^e\}_{\tau>0}$. In practice, two main data limitations need confronting, reflected in our subsequent empirical analysis: First, it is not possible to get a handle on the full sequence of the relevant effort and ability inputs that students have received since the start of their formal schooling. As a consequence, while we do have a strategy for distinguishing contemporaneous teacher ability and effort in 2002-03 (using the NCLB shock), our approach will be to summarize inputs from the more distant past using the lagged test score of a given student. Second, and related, identifying the full sequence of persistence parameters is not feasible, so we will concentrate on identifying a subset.[14]

Our proposed estimation strategies (developed in the next two sections) will be built around a re-writing of the test score technology in terms of once-lagged test scores,[15] allowing for the teacher j' and school s' in the previous year to be different, yielding:

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

The resulting expression for test scores, having moved the lagged score back to the RHS and with the relevant relabeling, is then

$$y_{ijgst} = \gamma y_{i,j',g-1,s',t-1} + a_{j(i,t)} + e_{j(i,t)} + \epsilon_{ijgst}\,. \tag{3}$$

[14] For the parameters governing the persistence of teacher ability, we will use strategies developed in prior work (CFR 2014b) to estimate the dynamic effects of teacher ability up to four periods in the past, given by $\gamma_1^a, \gamma_2^a, \gamma_3^a$, and γ_4^a. Estimating the persistent effects of effort is considerably more challenging, so as a first step, we focus on identifying the effect of once-lagged teacher effort, given by γ_1^e.

[15] This is accomplished by first multiplying 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 (see Todd and Wolpin 2003) – a composite measure of the persistent effects of teacher ability, teacher effort, and random shocks to performance. Second, subtract the result from both sides of the test score equation. (While we use a common persistence parameter γ for all prior accumulated knowledge as a starting point, our strategy will allow us to differentiate between persistence rates of ability and effort.)

where the error term contains the entire history of past inputs as well as the two most recent random shocks to performance.[16]

The first empirical goal of the paper can be stated with reference to equation (3): to separate out contemporaneous effort, $e_{j(i,t)}$ and ability, $a_{j(i,t)}$. That equation also guides our approach (presented in Section V) for estimating contemporaneous teacher ability and effort starting from teacher VA. Our empirical strategy will account for the potential endogeneity problem arising when students may be sorted to teachers based on their input histories (ϵ_{ijgst}), drawing on the introduction of NCLB, an intuitive account of teacher effort setting under NCLB, and standard estimation methods in the teacher VA literature.

Our second estimation goal, having recovered contemporaneous teacher ability and effort, relates to the *persistence* of the two input sequences. To isolate the persistence parameters that can be credibly identified, we re-write equation (3) and bring the once-lagged ability and effort terms – that is, $(\gamma_1^a - \gamma\gamma_0^a)a_{j(i,t-1)} + (\gamma_1^e - \gamma\gamma_0^e)e_{j(i,t-1)}$ – out of the error term ϵ_{ijgst}. Recalling our normalization that $\gamma_0^a = \gamma_0^e = 1$ and writing the resulting new error term as η_{ijgst}, the production technology can be expressed as

$$\begin{aligned} y_{ijgst} = \gamma(&y_{i,j',g-1,s',t-1} - a_{j(i,t-1)} - e_{j(i,t-1)}) \\ &+ a_{j(i,t)} + e_{j(i,t)} + \gamma_1^a a_{j(i,t-1)} + \gamma_1^e e_{j(i,t-1)} + \eta_{ijgst}\,. \end{aligned} \tag{4}$$

We can state the second goal with reference to equation (4): to identify the persistent effects of teacher ability and effort, given by γ_1^a and γ_1^e. This specification serves as the basis for the main estimation approach we implement in Section VI.

V Separating Contemporaneous Ability and Effort

In this section, we describe our estimation approach for separating a teacher's contribution to her students' test scores (measured by VA) into two inputs – *current* ability and effort – before presenting the estimates.

[16] As is clear from (2), when teacher ability and effort both decay at a common rate, γ, the error term consists of only random performance shocks, given by $\nu_{igjst} - \gamma\nu_{i,j,g-1,j,s',t-1}$. Otherwise, the error contains the entire history of inputs, given by $\sum_{0\leq\tau\leq g-1}\left[(\gamma_\tau^a - \gamma\gamma_{\tau-1}^a)a_{j(i,t-\tau)} + (\gamma_\tau^e - \gamma\gamma_{\tau-1}^e)e_{j(i,t-\tau)}\right] + (\nu_{igjst} - \gamma\nu_{i,j,g-1,j,s',t-1})$.

V.A Estimation Approach

We first provide an overview of our approach, which consists of three steps.

Step 1 – Estimating Teacher-Year Fixed Effects: We compute teacher-year fixed effects ($\hat{q}_{jt}$) for each 'teacher j and academic year t' combination using standard methods. Appendix B describes the students and teachers in the VA estimation sample and the approach we follow for computing teacher-year fixed effects.

As a precursor to the next two steps, we aggregate the assumed student production technology given by equation (3) up to the teacher level. This allows us to write the teacher fixed effect ($\hat{q}_{jt}$) as the sum of incentive-invariant teacher ability (a_j), incentive-varying teacher effort averaged across students in the class ($\bar{e}_{jt}$), and a common classroom shock that includes mean test score noise ($\bar{\epsilon}_{jt}$):

$$\hat{q}_{jt} = a_j + 1(t \geq \text{2002-03})\bar{e}_{jt} + \bar{\epsilon}_{jt}. \tag{5}$$

We separately identify the ability and effort components in equation (5) in the two steps that follow.

Prior to that, two comments about the RHS of equation (5) are in order. First, our estimates of teacher 'ability' should be thought of as potentially capturing both (true) ability and the average ABCs-related effort exerted by the teacher across all of her years of teaching under the ABCs program, given it operated in North Carolina prior to NCLB (see Appendix B.III for further discussion): the two cannot be separately identified. Second, understanding 'ability' in that sense, the equation makes the timing of the effort impact of NCLB explicit: the indicator variable multiplying average effort turns on when the academic year is 2002-03 or later. This timing will be key to our strategy for identifying variation in the effort component of the teacher fixed effect separately from the ability component.

Step 2 – Estimating Incentive-Invariant Ability: As our second step, we use pre-reform data to identify teacher ability during a period when NCLB did not operate, thereby ensuring that our estimates of teacher ability are independent of performance variation due to NCLB incentives. We do so using the Empirical Bayes (EB) estimator of teacher VA (see Kane and Staiger 2008, and Chetty, Friedman and Rockoff 2011), assuming incentive-invariant ability is

fixed over time, conditional on teacher experience.[17] Specifically, we estimate teacher ability by running the following pooled regression across all grades and years from 1996-97 to 2001-02, in which test scores are regressed on grade-specific cubic polynomials of prior scores (written $f_g(y_{i,j',g-1,s',t})$), indicators for student ethnicity, gender, limited-English proficiency, disability status, parental education, grade repetition, grade and year fixed effects, and controls for teacher experience:[18]

$$y_{ijgst} = f_g(y_{i,j',g-1,s',t}) + x'_{ijgst}\beta + h(exp_{jt}) + a_j + \theta_{jt} + \epsilon_{ijgst}, \tag{6}$$

where a_j represents teacher ability. 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, a_j, from classroom-specific shocks, θ_{jt}, and student-level noise, ϵ_{ijgst}.

Step 3 – Estimating NCLB-Induced Effort Response: In the third step, we estimate NCLB-induced teacher effort. We do so using the estimated teacher fixed effects from 2002-03 along with estimates of teacher ability and the fraction of students in a teacher's classroom deemed 'marginal' with respect to the NCLB target, drawing on the intuitive notion (already rehearsed) that teachers have the strongest incentives to direct additional effort to students predicted to score close to the proficiency threshold.

As in the descriptive analysis in Section III above, we define a student as 'marginal' if she is predicted to score within four developmental scale points of the test score proficiency cutoff on either side, though we consider various alternative definitions below. 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 constant

[17] For an estimator that allows teacher ability to drift over time, see CFR (2014a) and the studies of Rothstein (2014) and Bacher-Hicks *et al.* (2014). We opt not to use such an estimator, given that predicting teacher ability in 2002-03 with the drift estimator requires performance data from that year in the construction of optimal weights, and could confound teacher ability estimates with incentive variation in that year. We also note that the main advantage of the drift estimator is that it assigns greater weight to data from more recent years in order to better predict teacher performance in a given year. CFR (2014a) show this improves teacher VA prediction in their large urban unnamed school district. In North Carolina, the correlation between teacher effect measures over time is higher than in the CFR setting (see Rothstein 2014), implying a smaller benefit to using the drift estimator in our setting.

[18] We parameterize 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 CFR (2014a).

teacher incentive-invariant ability ($\hat{a}_j$) and teacher experience (exp_{jt}):

$$\hat{q}_{jt} = \psi m_{jt} + \lambda \hat{a}_j \ + w(exp_{jt}) + \xi_{jt}. \tag{7}$$

Once NCLB pressure is introduced, teachers may exert additional effort according to the amount of incentive pressure they face, and none before. We thus test whether there is a systematic relationship between $\hat{q}_{jt}$ and m_{jt} in 2002-03 but no relationship prior. Because we assume that teacher inputs contribute to overall teacher performance in an additive way, we use equation (7) to predict the portion of teacher performance given by effort, which we denote $e(m_{j,02-03}) \equiv \hat{\psi} m_{j,02-03}$.[19] This predicted value represents the response to the new incentive scheme, capturing the relationship between teacher-year performance and the classroom fraction of marginal students in 2002-03 (conditional on ability and experience).

V.B Results from Separating out Ability and Effort

We now present the resulting estimates for ability and effort.[20]

V.B.1 Estimating Ability and Effort

We start with our main estimates for contemporaneous teacher ability. Figure 3 presents the incentive-invariant teacher ability distribution, where incentive-invariant ability is defined as the EB estimate from equation (6). It indicates that there is significant variation in ability across teachers, all centered roughly around zero (due to a normalization in the EB procedure). Table 2 reports various summary statistics, including estimated standard deviations of 2.16, 1.63 and 1.63 developmental scale points across third, fourth, and fifth grade, respectively (see columns (1)-(3)). Averaged across grades, the standard deviation is 1.79 scale points, or equivalently 0.18 student-level standard deviations, which is within the range found by most previous studies.

Next, we present the estimated relationships between value-added and incentives – key

[19] This simple parameterization relating teacher performance to the fraction of marginal students is supported by the motivating visual evidence – see the binned scatter plots of Figure 2. The linear relationship, looking ahead, is also borne out in our results below (see Figures 4 and 6). Omitting an intercept, this parameterization relies on the assumption that NCLB-related teacher effort is zero when a teacher has no marginal students in her classroom. This serves as a reasonable approximation and does not affect any of our results below, which focus on the variance of predicted effort and marginal changes in effort, both of which are determined solely by the slope parameter, $\hat{\psi}$.

[20] Summary statistics for teacher-year fixed effects are reported in Appendix B rather than the main text, given that they are used solely as inputs for Steps 2 and 3.

findings from the analysis. The panels of Figure 4 show the grade-specific partial relationships between $\hat{q}_{jt}$ and m_{jt}, where the latter is residualized with respect to teacher ability and experience. For each grade, we plot the relationship for 2002-03. In 2002-03, 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. We also plot the corresponding pooled relationship for all pre-NCLB years that also includes year fixed effects. In contrast, the pre-NCLB variation plots reveal no discernible link between teacher performance and our measure of NCLB incentives prior to NCLB being in effect. These results are robust to alternative cutoff points for defining a student as marginal under NCLB (see Appendix C).

Regression estimates indicating how teacher-year effects change with an increase in the fraction of marginal students in the classroom are shown in Panel (a) of Table 3. This table reports the underlying estimates of ψ from equation (7). They 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 9 percent, 22 percent, and 16 percent standard deviation increases in teacher-year VA in third, fourth, and fifth grades, respectively. As expected, conditioning on teacher ability and experience, there is virtually no relationship between teacher-year effects and the classroom proportion of marginal students in the pre-NCLB years.[21]

The panels of Figure 5 present the full distributions of predicted effort in each grade in 2002-03, where effort is constructed as the fitted value $e(m_{j02-03}) = \hat{\psi} m_{j02-03}$. (Columns (7) to (9) of Table 2 present the corresponding estimates in each grade.) Mean teacher effort averaged across all grades is 0.61 points. Although the dispersion in teacher effort is not as high as the dispersion in teacher ability, we find quantitatively significant variation in effort across teachers: the variance of effort across all grades is 0.48 scale points, which equates to 0.05 student-level standard deviations of the test score.

[21] Here, one may worry about a mechanical correlation between teacher-year fixed effects and teacher ability, as the latter is estimated using pre-NCLB variation – the same variation as used to estimate pre-reform fixed effects. We address this problem by using jack-knife EB estimates of teacher ability in the pre-NCLB period, which use information from all *other* years excepting the one in question (Chetty, Friedman and Rockoff 2011).

V.B.2 Within-Teacher Performance Improvements

We now show that NCLB incentives caused performance improvements *within-teacher*. To that end, we construct the difference between 2002-03 and 2001-02 teacher-year fixed effects:

$$\begin{aligned}\hat{q}_{j02-03} - \hat{q}_{j01-02} &= a_j + e_{j02-03} + \bar{\epsilon}_{j02-03} - (a_j + e_{j01-02} + \bar{\epsilon}_{j01-02}) \\ &= e_{j02-03} - e_{j01-02} + \bar{\epsilon}_{j02-03} - \bar{\epsilon}_{j01-02}.\end{aligned} \tag{8}$$

On the RHS, these are written in terms of their ability and average effort components, which simplify to differences in effort and noise. We then regress the difference in teacher-year fixed effects on the fraction of marginal students faced by each teacher in 2002-03 ($m_{j,02-03}$) to explore whether stronger NCLB incentives caused a greater within-teacher performance improvements. If teachers with high fractions of marginal students 2002-03 were 'unlucky' in 2001-02 and had performed unusually poorly in that year, we would expect their performance to mechanically improve from one year to the next independently of the new NCLB performance incentives. To account for mean reversion, we control for a cubic function of 2001-02 teacher-year value-added. Specifically, the estimating equation is

$$\hat{q}_{j02-03} - \hat{q}_{j01-02} = \alpha + \chi m_{j02-03} + g(\hat{q}_{j01-02}) + \zeta_{j02-03}, \tag{9}$$

where χ is the main parameter of interest, reflecting any relationship between NCLB incentives and the within-teacher performance improvement, and $g(\hat{q}_{j01-02})$ is a cubic function of 2001-02 teacher-year VA.

The panels of Figure 6 show the partial relationships between the performance improvement in 2002-03 and $m_{j,02-03}$, while panel (b) of Table 3 reports the underlying slope coefficients (i.e., estimates of χ). Within-teacher performance improvements are clearly increasing in the fraction of marginal students in the classroom in 2002-03. A pooled regression of all pre-NCLB years (with transitions from year $t-1$ to t) is used as a placebo control in each grade, revealing a relatively flat relationship, and supporting the claim that the 2002-03 patterns reflect NCLB effort incentives.

V.B.3 Rival Hypotheses to Effort Setting

This evidence is consistent with teachers increasing effort in response to the incentives introduced under NCLB. Given that we do not observe effort directly, it is important to consider alternative hypotheses. One rival explanation is that students were sorted *differentially* to teachers in 2002-03, with high (incentive-invariant) ability teachers receiving greater fractions of marginal students. A second possibility is that schools might sort marginal students differentially into smaller sized classrooms in response to NCLB. We consider both hypotheses in Appendix C.II, conducting formal tests to assess how important each is. We demonstrate there that students are not sorted differentially to teachers (based on teacher ability) in a way that could explain our results, and that accounting for class size in our analysis does not change any of our main estimates.

VI Estimating the Persistence of Teacher Ability and Effort

We now assess whether teacher ability and effort persist at different rates. The issue is not just important for the policy analysis that follows: it is also key to understanding whether the separate effort effect we identify is likely to be consequential for economic outcomes in the longer run. As in the prior literature that estimates the persistence of teacher effects, we conduct the analysis at the student level (rather than the teacher level – the level of analysis in the previous section).

VI.A Estimating the Persistence of Ability

We estimate the persistence of ability in a reduced-form way, following the previous literature (see CFR 2014b, for example). Specifically, we regress student test scores in academic year $t+n$ (where n ranges from -2 to 4) on the full control vector from the Empirical Bayes regression (equation (6)) and the ability of teacher j who taught the student in period t:

$$y_{i,j,g,s,t+n} = f_g(y_{i,j',g-1,s',t}) + x'_{ijgst}\beta + h(exp_{jt}) + \phi_n \hat{a_j} + \epsilon_{ijgst}. \tag{10}$$

Here, ability is measured with a jack-knife EB estimator, which uses information from all years except the current year to form the teacher ability estimate. Doing so avoids a mechanical correlation between measurement error in test scores and teacher ability from confounding the

results (see CFR 2014b for further discussion). The coefficient ϕ_n represents the degree to which the effect of teacher ability from year t influences test scores in year $t+n$.

Figure 7 presents the estimated ϕ_n coefficients for regressions based on test scores exclusively from the pre-NCLB period. It shows that teachers do not affect their students' test scores in the years before they are matched with these students, as shown by the estimate at $t-2$; since we control for once-lagged test scores when estimating teacher ability, the coefficient at $t-1$ is identically zero. The estimates indicate that a one developmental scale point better-than-average teacher in year t improves student test scores by almost exactly one developmental scale point, on average.[22] The contemporaneous effect of teacher ability then fades away over time, as we estimate that 41 percent of the initial effect persists in period $t+1$, and only 20 percent remains by period $t+4$. These results align closely with the prior literature (see CFR 2014b).

VI.B Estimating the Persistence of Effort

Next we turn to estimating the persistence of teacher effort. Our approach is designed to address three challenges (already rehearsed in the Introduction): First, incentives to exert effort under NCLB are strongly correlated over time,[23] which makes it important to account for the effects of contemporaneous effort on scores to avoid overstating the persistence of lagged effort. Second, contemporaneous effort will depend on the persistence parameter we wish to estimate, given that educators make effort decisions based on expected student performance and predicted scores are determined in part by effort persistence. Third, we must account for induced changes in ABCs effort that arise from the introduction of NCLB in order to avoid confounding school-level ABCs-related improvements with student-level effort persistence.

The approach involves the following main components: estimating effort at the student level; focusing on effort persistence one year ahead; using the production technology to account for the three relevant effort components just referred to, in the process drawing attention to a key policy parameter used in our policy framework in Section VII; and an estimation routine to recover the parameters of interest using maximum likelihood. We now summarize each of these components in turn – a detailed description can be found in Appendix D – before presenting the estimates.

[22] The point estimate is 0.998 with a 95-percent confidence interval of (0.983, 1.015).

[23] Students who are marginal in one year tend to be marginal in the next, and so would be expected to receive higher effort as a consequence.

VI.B.1 Estimating Student-Level Effort

The first key component involves constructing a student-level measure of effort (rather than one that is common to all students taught by a given teacher). Doing so allows us to exploit more variation in the data, as the proficiency-count design of NCLB implies effort incentives can vary across students *within* a given classroom – in 2002-03, for example, fully 75 percent of the variance in our incentive strength measure (defined below) occurs within-classroom. The student-level effort measure we construct draws on the non-parametric patterns in Figure 1, which show that the introduction of NCLB had pronounced non-linear effects on student test scores, consistent with strong teacher effort responses to the scheme.[24]

Our interpretation of Figure 1 depends on two key concepts (formally defined in Appendix D.II): First, we form a *predicted student score* for 2002-03 – one that does not include the NCLB-induced effort response. Taking the difference between the realized and predicted score for a given student then provides a (noisy) measure of the 2002-03 effort response by her teacher (where effort is taken as the incentive-related boost to scores). Second, we construct a measure of *incentive strength* for each student, which depends on the distance between the predicted score and the fixed NCLB proficiency target. This measure is used on the horizontal axis in Figure 1, which groups students into two-scale-point width bins of incentive strength in 2002-03 (denoted $\pi_{i,02-03}$). On the vertical axis, we then plot the *average* difference between the realized and predicted score (discussed above) within each incentive strength bin, eliminating idiosyncratic test score noise to recover average teacher effort as a function of incentive strength.

The pattern for 2002-03 shows that students who are predicted to score near the proficiency threshold ($\pi_{i,02-03} \approx 0$) – namely those for whom effort incentives are strongest – receive the largest boost to their scores. We conduct the same exercise for the 1999-2000 pre-reform period (when there can be no NCLB effort response) to ensure that we do not systematically under- or over-predict test scores for certain parts of the distribution. Doing so makes clear that our predicted score tracks the realized score very well throughout the distribution, given by the flat line. In turn, this lends credence to the view that the 2002-03 pattern reflects student-specific NCLB effort.

We use the profiles for the two years in Figure 1 to estimate a student-specific effort function

[24] Macartney *et al.* (2015) use similar patterns as the basis for a structural approach to study the design of incentives in education.

that takes incentive strength as its argument.[25] The resulting effort function, denoted by $e^N(\cdot)$, is plotted in Figure 8. We then use this function to assign a level of effort to each student. Taking the student-specific values of $\pi_{i,02-03}$ and the function $e^N(\cdot)$, the effort directed to each student i in 2002-03 is given by $e_{j(i,02-03)} = e^N(\pi_{i,02-03})$, reading off from the function.[26]

VI.B.2 The Components of the Estimating Equation

With the student-specific effort measure in hand, we specify an equation that can be taken to the data in order to estimate the rate at which such effort persists. Here, we draw on the technology presented in the conceptual framework, given by equation (4). That equation allows us to express an individual student's score in 2003-04 in terms of the persistent effect of once-lagged scores from 2002-03 excluding teacher ability or effort, the effects of teacher ability and teacher effort in the current year 2003-04, the persistent effects of teacher ability and effort from 2002-03, and a random shock to current test scores. Effort in 2003-04 can then be further subdivided into two parts: (i) effort arising from contemporaneous NCLB incentives, and (ii) effort that arises from altered school-level ABCs incentives as a consequence of the NCLB effort response in the previous year (2002-03). The reasoning is set out in full in Appendix D.III.

Formally, the main estimating equation in 2003-04 is given by:

$$y_{i,j,g,s,03-04} - y^C_{i,j,g,s,03-04} = \gamma^e_1 e^N(\pi_{i,02-03}) + \theta e^N(\pi_{i,03-04}) + \rho \bar{e}^N_{s,02-03} + \eta_{i,j,g,s,03-04}. \quad (11)$$

We describe each component in turn,[27] starting on the LHS. This consists of the difference between the actual test score of student i in 2003-04 ($y_{i,j,g,s,03-04}$) and what we refer to as the 'counterfactual predicted score,' denoted $y^C_{i,j,g,s,03-04}$. The latter represents the test score student i *would* have earned in 2003-04 had NCLB not been enacted in the prior year (in which case there would be no contemporaneous or persistent effect of effort).[28] Deducting the counterfactual

[25] Specifically, we difference the binned 2002-03 and 1999-00 profiles, then fit an eighth-order polynomial to the differenced data using a weighted regression, with the weights capturing the total number of students in each bin (across both 2002-03 and 1999-00).

[26] Note that the values of the effort function are negative for the left and right extremes of incentive strength. We interpret the effort function as reflecting student test score gains *relative* to the pre-NCLB status quo. In that sense, the extremes of incentive strength are not associated with negative levels of *absolute* effort but with lower test score gains than in the pre-NCLB period for select non-marginal students.

[27] For reference, a complete listing of all the formal notation used in this section is given in Appendix Table D.1.

[28] It is calculated by applying the predicted score procedure from the prior subsection a second time, now using the predicted score from 2002-03 as an input.

from actual score on the LHS cancels out all non-effort inputs on the RHS, allowing us to focus exclusively on the effort components in 2003-04 that are relevant from an estimation perspective.

The first term on the RHS – effort in 2002-03 – is known by the econometrician. We assume it is determined by incentive strength ($\pi_{i,02-03}$), according to the semi-parametric effort function $e^N(\cdot)$, which reflects NCLB's introduction in the previous year. The parameter γ_1^e captures the rate at which effort in 2002-03 persists to affect test scores in 2003-04.

The second term, unknown to the econometrician, consists of NCLB-induced effort in 2003-04. This is computed based on the plausible assumption that the effort devoted to student i in 2003-04 is given by $\theta e^N(\cdot)$ evaluated at $\pi_{i,03-04}$, where $\theta > 0$. Thus, teachers use the same empirically-determined effort function as in 2002-03 to set effort, with the given function taking $\pi_{i,03-04}$ as its argument in 2003-04, and the parameter θ either diminishing (when $\theta < 1$) or amplifying (when $\theta > 1$) all effort levels in a proportional way. By way of justification, it is plausible to think that teachers would direct effort to students in a similar fashion across the two years, with marginal students receiving relatively more effort than non-marginal students in each year, given that the rules of NCLB remained constant across 2002-03 and 2003-04. We assume that teachers predict each student's incentive strength in 2003-04 ($\pi_{i,03-04}$) based on the student's counterfactual predicted score ($y^C_{i,j,g,s,03-04}$), which does not include a prior effort response, and the amount of effort that persists from the previous year ($\gamma_1^e e^N(\pi_{i,02-03})$).[29]

The third term accounts for the effects of ABCs effort on student test scores in 2003-04, depending on average school-level effort from 2002-03 according to the parameter ρ. This term captures the fact that NCLB effort responses in 2002-03 affect school ABCs performance targets in 2003-04, thereby also potentially affecting school effort decisions under the ABCs. Because school ABCs targets are functions of student *average* prior-year test scores, they also depend on *average* prior-year effort (see Appendix D.III). Thus, changes in average school-level effort from 2002-03 lead indirectly to changes in schools' ABCs effort decisions in 2003-04 (through the effect of prior effort on the subsequent ABCs targets and passing probabilities of schools). The parameter ρ therefore reflects the indirect effect of changes in ABCs incentives on student test scores. We also note that the exclusive use of school-level growth targets under the ABCs, without corresponding student-level targets, implies that incentives to exert effort do not vary

[29] We assume that teachers make predictions about student performance with full knowledge of both the true value of γ_1^e and the level of effort each student received in the prior year (as detailed in Appendix D.III).

across students within a given school.[30]

VI.B.3 Estimation Approach and Identification

Our goal is to recover the parameters of equation (11), as described in Appendix D.III. While our primary focus is on γ_1^e, which governs the persistence of effort, we also pay close attention to the estimate of ρ, as it is key to our subsequent policy calculations in Section VII. As previously mentioned, the estimation challenge is that the input to the 2003-04 effort function depends on the (unknown) persistence rate. Yet in order to estimate the persistence rate, we need to account for the correlation of effort across time. We address this challenge by using maximum likelihood to recover γ_1^e, effort in 2003-04 (given by $e^N(\pi_{i,03-04})$), and the scale factor modifying the 2003-04 effort function (θ), making a normality assumption about the error in equation (11). Our maximum likelihood procedure is described more fully in Appendix D.IV.

As we now discuss, the parameters γ_1^e, θ and ρ, are all separately identified. In particular, the identification argument for the parameters associated with NCLB incentives (γ_1^e and θ) does not depend on the parameter associated with ABCs incentives (ρ). Further, the latter is separately identified from both γ_1^e and θ.

Identification of NCLB Parameters (γ_1^e, θ). Separate identification of γ_1^e and θ requires that, conditional on 2002-03 effort, given by $e^N(\pi_{i,02-03})$, there is remaining variation in 2003-04 effort ($e^N(\pi_{i,03-04})$) and vice-versa. Such variation is guaranteed by the non-monotonic shape of the effort function in 2002-03, which ensures that two students with the same level of effort in 2002-03 can have different levels of NCLB incentive strength and, correspondingly, different levels of 2003-04 NCLB effort.[31] Given the 2002-03 effort function is non-monotonic, identification of the parameters of interest requires a minimal condition: the 2003-04 effort function should not be flat.[32]

[30] This assumption is testable. Given that NCLB provides differential effort incentives across students, the estimates of the NCLB effort parameters, γ_1^e and θ, should not change appreciably across specifications that do and do not control for lagged school-level effort. We provide relevant evidence below.

[31] To see that this is the case, refer to Figure 8 and consider two students, each with a 2002-03 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 11 scale points. These students will continue to have different predicted scores in 2003-04 and will then receive different levels of effort in that year. The student with prior incentive strength of -3 receives *more* effort in 2003-04 than she did in 2002-03. The student with prior incentive strength of 11 also gets a bump in his predicted score, which moves him up in the incentive strength distribution, but to a point where he receives *less* effort in 2003-04 because he is now further away from the proficiency threshold.

[32] The argument is set out in Appendix D.IV. Although we assume the same functional form for the effort function (up to the scale θ) in 2003-04 as in 2002-03, this assumption is not required for identification: any effort

Identification of ρ. To identify the separate effects of NCLB and ABCs incentives, the identifying assumption is that ABCs incentives operate *across* schools while NCLB incentives operate *within* schools. This is reasonable given that the ABCs scheme sets only an average school-level growth target, while NCLB sets a student-level target (the test score required for subject matter proficiency) in addition to an overall school-level target (the proficiency rate). Separate identification of ρ from both γ_1^e and θ relies on there being significant within-school variation in NCLB incentives, a condition satisfied in the data.

VI.B.4 Maximum Likelihood Estimates

Having established how the parameters are identified, we apply the maximum likelihood estimation routine to a sample of students who transition grades between 2002-03 and 2003-04, and for whom test scores are non-missing in each year. This restriction allows us to compare realized scores with counterfactual predicted scores for each student.

Table 4 presents the results. The first column provides an estimate of the persistence of NCLB effort from 2002-03 without accounting for contemporaneous NCLB effort in 2003-04. In this case, 40 percent of the initial effort effect persists one year into the future: as shown in column (2), this estimate overstates the persistence rate. Once we account for *contemporaneous* NCLB effort, the estimate of γ_1^e falls to 0.10, implying that only 10 percent of the initial effort persists to affect 2003-04 test scores. The estimate of $\theta = 0.52$ in column (2) indicates that the effort response is scaled down by around 50 percent in 2003-04 relative to 2002-03. This finding of $\hat{\theta} < 1$ implies that the difference between the effort received by marginal and non-marginal students at the average school becomes smaller in 2003-04 than in 2002-03, suggesting a lower relative boost to marginal student test scores over time.

Accounting for ABCs incentives in column (3) yields an estimate of $\hat{\rho} = 0.29$, meaning that a one standard deviation increase in school-level NCLB effort from 2002-03 ($\bar{e}^N_{s,02-03}$) produces a 0.8 percent of a standard deviation increase in student-level test scores.[33] While this is a relatively small effect, the positive and significant estimate of ρ is consistent with the 2002-

function in 2003-04 that is not flat would be sufficient. (The functional form we do use for the effort function is not assumed but rather is *estimated*, based on our strategy above treating the introduction of NCLB in 2002-03 as an exogenous shock to incentives.)

[33]The standard deviation of $\bar{e}^N_{s,02-03}$ across schools in 2003-04 is 0.27 developmental scale points. Multiplying 0.27 by $\hat{\rho} = 0.29$ gives an effect of 0.078 developmental scale points, or 0.8 percent of a student-level standard deviation.

03 NCLB effort response strengthening ABCs incentives in the following year by making the targets more difficult to pass, in turn leading to student performance gains (as discussed in Section VII.A). The estimates of γ_1^e and θ are nearly identical to those obtained from the maximum likelihood routine that does not account for ABCs incentives, lending credence to the notion that NCLB incentives vary *within* schools while ABCs incentives vary *across* schools (as the respective accountability incentives imply).

VII Framework for Policy Analysis

Building on the conceptual model and estimates above, this section develops a framework for measuring the cost effectiveness of incentive-based education reforms, allowing them to be placed alongside feasible alternatives. The general approach involves comparing pairs of policies – one incentive-based and one not – in terms of their costs and benefits.

Taking these in turn, it is not always possible to express the *costs* of a given policy in units that are common across reforms (a natural candidate being dollars). In particular, the incentives we draw on under NCLB to elicit a given amount of extra teacher effort arise from a set of non-pecuniary sanctions; they are specific to that policy implementation and are not readily translatable to non-incentive-based policies. To overcome that hurdle, we devise a procedure (explained in Section VII.A) to estimate the value of the sanctions in dollars using our conceptual model and features of the NCLB and ABCs reforms, creating a bridge between the two incentive programs.[34]

The benefit side of the comparison is more straightforward, as benefits can be well defined in terms of student achievement. We use our estimates in order to set the incentives under the incentive-based policy to equalize the implied student performance effects across the two policies (see Section VII.B), a task the estimates of the contemporaneous and persistent effects of teacher effort are essential for. Having monetized the sanction and equated the benefits of the two reforms, we are then able to compare the cost of the NCLB incentive reform with alternatives.

This type of cost-effectiveness comparison involving incentive-based reforms is new to the literature. We illustrate the approach by comparing incentive-based policies for improving teacher

[34] In particular, we take advantage of the fact that ABCs incentives are pecuniary, and note that the initial NCLB effort response affected subsequent ABCs incentives and effort.

effectiveness with candidate policies that aim to improve average teacher productivity through the dismissal of low-performing teachers. Such 'ability-based' measures, already mentioned in the Introduction, have received considerable attention in recent work, with Hanushek (2009, 2011), CFR (2014b), and Rothstein (2015) all analyzing policies that involve the dismissal of teachers whose value-added falls in the bottom part of the measured distribution (for example, the bottom five percent).[35] Owing to the fact that such reforms are only effective for a subset of teachers, while incentive-based reforms can be designed to be applicable for all teachers, one must adjust the relative pecuniary cost of the reforms, which we do in Section VII.C.

VII.A The Cost of Each Reform

We begin by determining the costs of the ability-based and incentive-based reforms, specifically focusing on an ability-based reform that involves dismissing the bottom five percent of teachers in the VA distribution. As is well-appreciated, this type of reform creates increased employment risk for teachers throughout the distribution due to estimation error in value-added measures. Rothstein (2015) estimates that compensating teachers for the increased risk would require 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. Thus, we assume that implementing the ability-based reform in our setting comes at an additional cost of $700 per teacher.

Isolating the costs of the incentive-based reform requires an estimate of the per-teacher monetary equivalent of the NCLB sanction used in practice. We offer a quick overview of the approach we devise here, before breaking it into four distinct steps.[36] Our approach begins by estimating the effort response to incentives under the ABCs – a nontrivial task since such incentive variation is endogenous (responses to NCLB affect incentives under the ABCs). Our solution is to exploit the dynamic interaction between the NCLB effort response in 2002-03 and the ABCs effort response in 2003-04, as the former response made subsequent ABCs targets more difficult to satisfy. In particular, we are able to quantify both the expected financial loss and the increased effort response under the ABCs as a result of the prior exogenous NCLB shock. Combining these two quantities allows us to identify the link between monetary rewards

[35] The literature has also focused on reducing the attrition of the highest rated teachers. Existing research (CFR 2014b) suggests that a focus on the top is a less cost-effective ability-oriented reform than replacing the lowest rated teachers.

[36] Additional detail is provided in Appendix E.I.

and effort under the ABCs. Assuming that the incentive-effort connection under the NCLB and ABCs reforms is the same, we then use this link to determine the strength of financial incentives (measured in dollar terms) that would give rise to the 2002-03 NCLB effort response observed in the data, thereby monetizing the sanction. The four steps are as follows:

Step 1– NCLB's Effect on Subsequent ABCs Financial Incentives: We calculate the degree to which school responses to NCLB in 2002-03 lowered the probability of passing the ABCs in 2003-04, relative to the probability in the counterfactual scenario in which NCLB was not enacted. The differences in these passing probabilities (in combination with the ABCs bonus payment of $750, received by teachers when their school satisfied its growth targets) determine the expected dollar value each school stood to lose because of its response to NCLB's introduction.[37] We obtain the corresponding change in 2003-04 financial incentives for a one-unit change in average school-level effort by regressing the expected dollar value each school stood to lose in 2003-04 on average school-level effort in the prior year.[38]

Step 2 – NCLB's Effect on Subsequent Student Test Scores: We recover the parameter ρ (estimated in Section VI), which measures the effect of lagged school-level effort on test scores in 2003-04. It represents the *indirect* effect of ABCs financial incentives (which NCLB alters – see Step 1) on test scores due to changes in teacher effort. This follows from 2002-03 school-level effort affecting schools' effort incentives in 2003-04 exclusively through the change it causes to the likelihood of a school passing the ABCs: conditional on student-specific effort from 2002-03, it should not have any direct effect on student test scores in 2003-04.

Step 3 – Connecting Changes in ABCs Financial Incentives and Test Scores: We estimate the *direct* effect of ABCs financial incentives on teacher effort by combining the effect of lagged effort on subsequent financial incentives (from Step 1) and on subsequent test scores (from Step 2). This step is analogous to the two-stage least squares procedure, in which we have estimated a first-stage effect (Step 1) and reduced-form effect (Step 2), and would like to obtain

[37] We make these calculations using the rules of the ABCs program and our conceptual framework.

[38] The resulting estimate governs the magnitude by which a one-unit increase in school-level effort in 2002-03 lowers the likelihood of ABCs target attainment in 2003-04. Panel (b) of Appendix Table E.1 reports the estimated coefficients from these regressions, the estimates implying that a one-unit increase in average 2002-03 school-level effort reduces the probability of passing the ABCs by between 9 and 32 percentage points. This range reflects a variety of plausible assumptions about the variance of the unobserved performance shocks to school-level ABCs growth scores (see Appendix E.I for more detail).

an estimate of the second-stage effect. Dividing the reduced-form estimate by its first-stage counterpart accomplishes this, recovering the effect of financial incentives on test scores due to changes in teacher effort.

Step 4 – Inferring the Value of the NCLB Sanction through the ABCs Link: The dynamic link between 2002-03 NCLB effort and the 2003-04 ABCs response allows us to estimate (in Steps 1 to 3 above) the responsiveness of teacher effort to financial incentives. We then use this incentive-effort relationship to infer the *monetary* value of the sanction that generated the observed NCLB effort response in 2002-03. Specifically, the average school-level NCLB effort gain in 2002-03 ($\bar{e}^N_{s,02-03}$) is 1.97 developmental scale points, suggesting that the NCLB sanction is valued between \$458 and \$1,640 per teacher, where the range reflects different plausible assumptions about the variance of the unobserved performance shocks to school-level ABCs growth scores (corresponding to the range in Step 1). In sum, based on the observed NCLB response, our upper-bound estimate for the cost of the NCLB sanction is therefore \$1,640.

VII.B Equating the Student Achievement Benefits of Each Reform

In order to carry out a cost-effectiveness comparison of the two reforms in the following subsection, we first place the two types of policy – ability- and incentive-based – on a common footing in terms of their effects on student achievement. CFR (2014b) provide benchmark estimates for ability-based reforms. Those indicate that replacing the lowest-rated teachers with random draws from the distribution of new teachers results in an average two standard deviation improvement in teacher ability (measured in student test scores units) for that subset. Although our sample differs somewhat, we take those values as estimates of the benefits of a similar ability-based policy in our setting.

Teacher productivity can be raised by an equivalent two standard deviations, on average, through an incentive-based reform by making two adjustments (see Appendix E.II for more detail). The first adjustment raises the fraction of marginal students in each classroom – up from the sample average of 26 percent – by introducing student-specific targets (instead of targets fixed within grades, as under NCLB). We estimate that making all students marginal would result in an effort response equivalent to 0.29 SD of the test score on average, corresponding to 1.6 SD of teacher ability. The second adjustment – needed in order to raise the response by the

required additional 0.4 SD of teacher ability – involves altering the monetary-equivalent value of the NCLB sanction. We show that a sanction set at 25 percent of its current value results in the desired performance increase. Here, the fact that the incentive reform affects *all* teachers while the ability-based reform only affects the lowest-performing subset allows us to set a monetary equivalent at significantly less than 100 percent of the current sanction value.[39]

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

In this subsection, we bring the estimates from the preceding subsections together to compare the cost of the two reforms when they produce the same benefit in terms of student achievement. Section VII.A suggests that a program providing a bonus payment of up to $1,640 per teacher when schools succeed would replicate the effort response, and thus the benefit in terms of test scores, observed under the sanctions-based NCLB. However, recall that our achievement estimates indicate the benefit under the actual NCLB reform would be smaller than the benefit under the leading ability-based reform (both in the present and persisting into the future).

Section VII.B places the benefits on a level playing field by making two adjustments, each of which has implications for the cost of the incentive-based reform. The first adjustment is to set student-specific targets to increase the fraction of marginal students across classrooms. While this adjustment may seem costless (aside from any administrative disruption), it actually results in substantial savings since the cost of the reform should be scaled down to compensate for the fact that the ability-based reform only affects a *subset* of teachers, as opposed to *all* teachers under the incentive reform.[40]

The second adjustment involves setting a higher monetary-equivalent value for the NCLB sanction and is therefore associated with new costs. However, these additional costs are far outweighed by the savings from the first adjustment: the cost of the incentive reform should be reduced by a *factor* of 20 under the first adjustment if the ability-based reform replaces the bottom five percent of teachers (a popular cutoff with policy makers), while the cost only rises

[39] Our calculations are based on the reduced-form estimates in Table 3 and plausible assumptions about the long-run persistence of effort (see Appendix E.II).

[40] The effect of the leading ability-based reform is always restricted to a subset of the distribution, since it entails firing the bottom 'X' 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 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. Eventually, one would face the prospect of dismissing teachers with above average value-added, which would yield negative returns. For this reason, ability-based policies offer a less flexible policy lever (compared to incentive-based policies) for improving average teacher performance.

by a quarter under the second adjustment. Overall, our proposed reform requires a per-teacher monetary equivalent of the NCLB sanction that is 25 percent of its current value, in order for the incentive- and ability-based policies to generate the same average increase in test scores.

Given the preceding discussion, the proposed incentive reform is associated with a cost of $410 (that is, $1,640*0.25) per teacher, or 59 percent of the $700 cost required to compensate teachers for increased employment risk under the popular ability-based reform. In other words, implementing the incentive-based reform involves compensating teachers around *40 percent* less than under the ability-based reform, despite the stronger achievement effects for ability that we estimate.

VII.D The Framework's Broader Applicability, Refinements and Extensions

Our framework allows incentive-based education reforms to be compared for the first time with a range of alternative policies that seek to raise school and teacher performance. In this way, it moves the policy conversation forward. The basic requirements are that the impact of such policies on student outcomes and their associated costs can be quantified.

The approach we propose is clearly not the final word on the subject – one can think of various ways in which the policy computations could be further refined. For example, our calculations use the approximation that teacher effort increases linearly in both the fraction of marginal students in their classrooms and the value of the NCLB sanction, while in practice, the effort responses might differ from those we estimate. We also draw estimates from prior work that may not accord with our setting. For instance, we use the 1.4 percent mean salary increase estimate from Rothstein (2015) to estimate the relevant compensation teachers would need for elevated employment risk, when a lower value would raise the attractiveness of the ability-based policy. Similarly, the two standard deviation improvement from replacing the lowest-rated teachers with random draws that CFR (2014b) estimate for a different setting may differ when applied to North Carolina teachers. Further research can help shed light on the role of the assumptions we make and the relevant variability in the estimates we use, in turn allowing us to assess how the policy comparison calculus in this paper might be affected.

In addition to those types of refinement, the basic approach we outline could be extended in various fruitful ways. First, one could envisage policies that were a weighted average of the separate incentive-based and ability-based policies that we compare. Second, one might

consider the extensive margin effect of ability-based and/or incentive-based reforms on the *stock* of teachers, as the threat of dismissal or sharpened incentives could affect teacher turnover via participation constraints. Third, contrary to our assumption, losses and gains may have asymmetric effects on teacher behavior, making reward-based incentive reforms more or less efficient than sanction-based reforms. Fourth, the use of extrinsic incentives could influence the intrinsic motivation of teachers, which would represent an additional cost of incentive-based policies. Fifth, the threat of dismissal under an ability-based policy could itself induce an effort response among teachers near the dismissal cutoff,[41] though it would likely be small compared to the response under an incentive-based policy: rather than applying to the full distribution of teachers, it would be confined to teachers within a narrow band around the cutoff.

Having noted these possible refinements and extensions, overall we view our analysis as a useful first step in bringing incentive-based and other education policies together, in the process showing that incentive reforms can be highly competitive with the alternatives. The implementation of our framework implies that the incentive reform can deliver the same benefit as the ability-based reform for one quarter of its current cost, making it more cost effective in the long run. That conclusion continues to hold even if the pecuniary value of the sanction were more expensive, up to an upper bound of 42 percent ($\$700/\$1,640$) of its current cost,[42] leaving room for assumptions to move against us. Even if they were to, our sense is that the effect of the incentive reform is likely to remain competitive with its ability-based counterpart, and that incentive reforms are worth considering seriously as a viable tool for education policy makers.

VIII Conclusion

Our primary goal in this paper has been to understand the way measured teacher performance is influenced by accountability incentives. To that end, we presented an approach permitting us to separate out teacher effort, which is responsive to education incentives, from teacher ability, which is not. Central to our approach was a novel identification strategy leveraging a

[41] Dee and Wyckoff (2015) explore such issues in their analysis of IMPACT, a teacher evaluation program in the District of Columbia. They show that the threat of dismissal both increased attrition and improved performance (for teachers who stayed in the profession) among teachers near the bottom of the performance distribution. IMPACT also featured a financial bonus for high-performing teachers, which improved performance and increased retention among teachers near the top of the performance distribution.

[42] This upper bound is obtained by dividing the per-teacher cost of the ability-based reform (which is required to compensate teachers for the additional employment risk) by our estimate of the per-teacher monetary-equivalent value of the NCLB sanction.

natural experiment associated with the introduction of a federal accountability program (NCLB) 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 linear relationship between teacher value-added and this incentive measure in the year NCLB was introduced but not in prior years. We then exploited these differences over time to separate teacher quality into teacher ability and the effort response associated with NCLB, allowing us to gauge the respective impacts of effort and ability on contemporaneous scores.

To measure the extent to which these two potentially important education inputs might persist differentially, we then developed a structural estimation strategy based on a transparent specification of the education production technology, which allowed us to identify the persistence of effort separately from the effects of contemporaneous incentives. Here, we found that effort persists at approximately 25 percent of the persistence of ability, the latter having a significant positive effect on future test scores. These estimates combined indicate that teacher effort is both a productive input and one that is responsive to incentive variation in a systematic way, with longer-term benefits for students.

Using the estimates and technology, we then explored the cost-effectiveness of incentive-based reforms alongside alternative education reforms discussed in the literature. Incentive-focused education policies have become increasingly widespread over the past two decades, yet how they compare with alternative types of education reform has remained under-explored, in part because of a lack of a framework for quantitative policy comparison. This paper proposed such a framework, allowing the cost effectiveness of feasible policies (including incentive-based reforms) to be compared for the first time. Our illustrative analysis indicated that using formal incentives constitutes a viable alternative way of accomplishing the goal of raising student and school performance – indeed, one that 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 refinements to the estimation approach we develop – for instance, looking at the longer-run persistence of effort. These are areas we are exploring in related research.

References

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.

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 James Wyckoff. 2015. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT." *Journal of Policy Analysis and Management.* 34(2): 267-297.

Deming David J., Cohodes Sarah, Jennings Jennifer, Jencks Christopher. 2016. "School Accountability, Postsecondary Attainment and Earnings." *Review of Economics and Statistics*, 98(5):848-862.

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

Hanushek, Eric A. 2009. "Teacher Deselection." in *Creating a New Teaching Profession*, ed. Dan Goldhaber and Jane Hannaway, 165-80. Washington, DC: Urban Institute Press.

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

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.

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.

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.

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.

Tables

Table 1: Student-Level Summary Statistics

	Mean	SD	Observations
Performance Measures			
Mathematics 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
Mathematics 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
Future[(a)] Mathematics 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
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
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
Future[(a)] 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
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 1996-97 to 2004-05.
[(a)] Future mathematics and reading scores are the scores we observe for our sample of third through fifth grade students when they are in sixth, seventh, and eight grades. 'Future' mathematics and reading scores are used when measuring the persistent effects of teacher ability and effort. We do not follow students past 2004-05, as the mathematics scale changes again in 2005-2006 yet no table to convert scores back to the old scale was created by the state. The free or reduced-price lunch eligibility variable is not available prior to 1998-99.

Table 2: Teacher Performance Variables

	Estimated Ability			Fraction of Marginal Students, m_{jt}			Estimated Effort		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Grade	3rd	4th	5th	3rd	4th	5th	3rd	4th	5th
Mean	-0.07	-0.09	-0.06	0.33	0.21	0.23	0.56	0.80	0.45
Observed SD	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	-	-	-	-	-	-
Observation	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 estimated ability are calculated using all teacher-grade observations from 1996-97 to 2001-02, 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 the fraction of marginal students in classrooms are calculated using all available teacher-year observations from 1996-97 to 2002-03. Because second grade scores are not available in 1996-97 and the change to the mathematics developmental scale in 2000-01, we are unable to calculate marginal status for third graders in 1996-97 and 2000-01, and for fourth and fifth graders in 2001-02. Summary statistics for estimated teacher effort are calculated across all teacher observations in 2002-03.

Table 3: The Effects of NCLB Incentives on Teacher Performance

	Panel (a): Teacher-Year VA as Dependent Variable					
	Third Grade		Fourth Grade		Fifth Grade	
	2002-03	Pre-NCLB	2002-03	Pre-NCLB	2002-03	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	
	2002-03	Pre-NCLB	2002-03	Pre-NCLB	2002-03	Pre-NCLB
Effect of m_{jt}	2.08***	0.18	4.11***	-0.16	2.48***	0.49***
	(0.18)	(0.16)	(0.31)	(0.18)	(0.24)	(0.17)
Observations	2,651	9,697	2,453	9,087	2,397	8,357

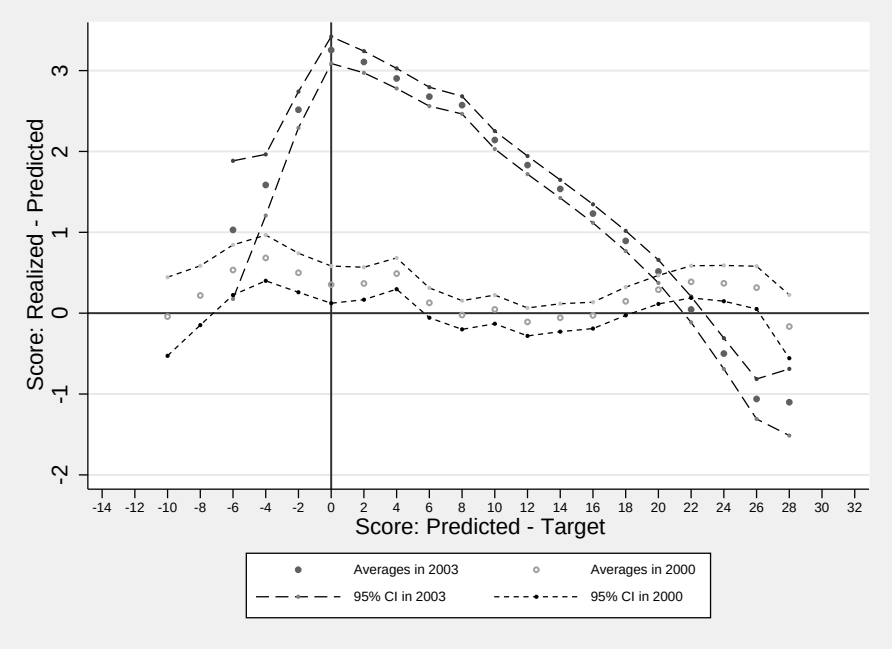
Notes: In panel (a), we present estimates of ψ from grade-specific regressions of equation (7). In the year 2002-03 regression, additional controls include teacher ability and teacher experience. The estimate in the pre-NCLB columns comes from a pooled regression of all pre-NCLB years that additionally includes year fixed effects. For third grade, the pre-NCLB years stretch from 1997 to 2000, and 2002; for fourth and fifth grade, they stretch from 1997 to 2001. In panel (b), we present estimates of χ from grade-specific regressions of equation (9). Specifically, we regress the change in teacher-year VA from 2001-02 to 2002-03 on the fraction of marginal students in the classroom in 2002-03 and a cubic function of 2001-02 teacher-year VA. We also regress the change in teacher-year VA from year $t-1$ to t in the pre-NCLB period (using a pooled regression of all years) on the fraction of marginal students in the classroom in year t, year fixed effects, and a cubic function of year $t-1$ 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.

Table 4: 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_1^e	0.40***	0.10***	0.10***
	(0.02)	(0.02)	(0.02)
θ	-	0.52***	0.49***
	-	(0.02)	(0.02)
ρ	-	-	0.29***
	-	-	(0.06)
μ	1.19***	0.90***	0.39***
	(0.04)	(0.04)	(0.12)
σ^2	20.30***	20.14***	20.14***
	(0.14)	(0.14)	(0.14)
Observations	86,236	86,236	86,236

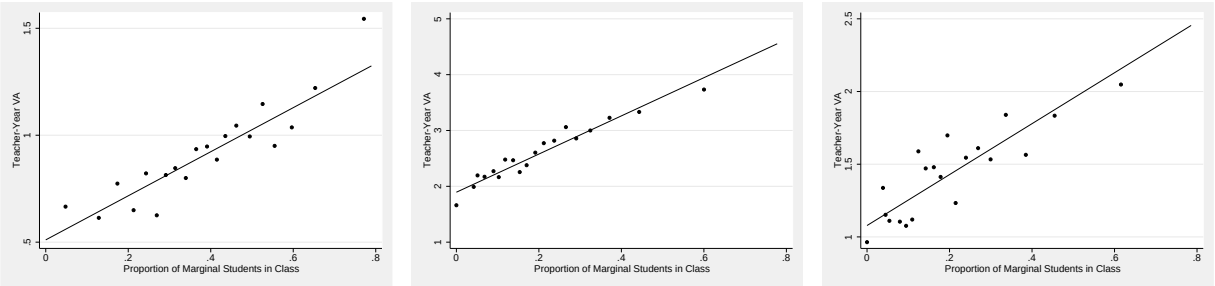
Notes: This table presents maximum likelihood estimates of variants of equation (11). The sample includes fourth grade students in 2003-04. The dependent variable in each column in the difference between the realized and counterfactual predicted mathematics score. Standard errors calculated using the Outer-Product of Gradients method appear in parentheses. *** denotes significance at the 1% level.

Figures



Notes: This figure shows the effect of accountability incentives on fourth grade mathematics scores. It is constructed as follows: In each year, we calculate a predicted score for each fourth grade 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 academic year 2002-03 averages; the hollow circles are academic year 1999-00 averages. The figure also shows the associated 95 percent confidence intervals for each year. Standard errors are clustered at the school level.

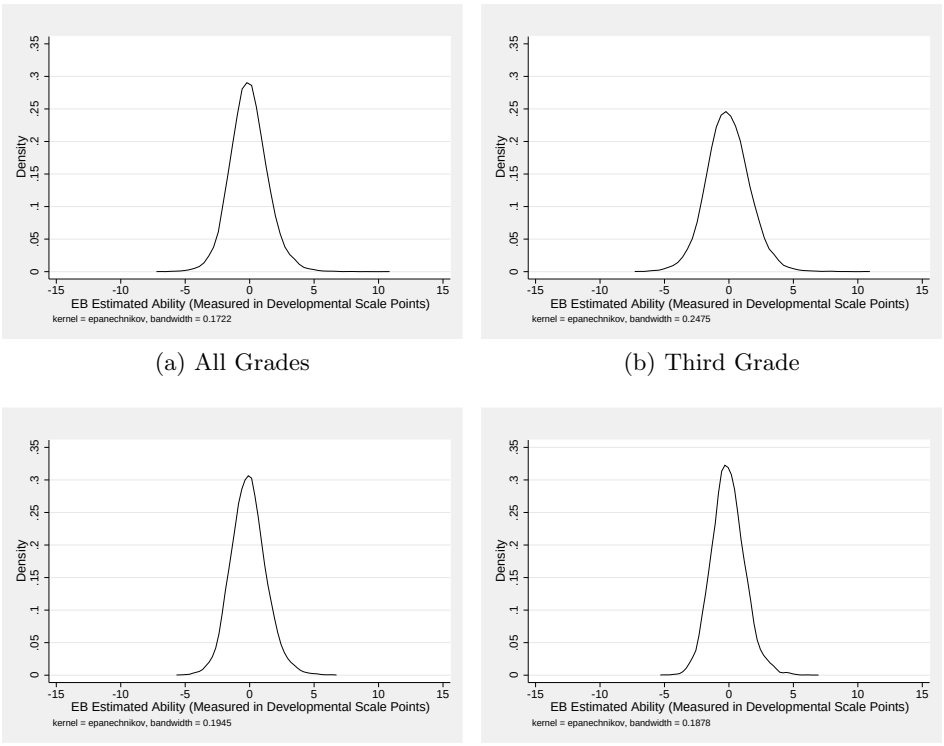
Figure 1: Inverted-U Response to NCLB's Introduction



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

Notes: The panels in this figure depict the relationship between teacher-year VA measures and the fraction of marginal students within a classroom in the academic year 2002-03. To construct the figure, we first group teacher-year observations into 20 equally-sized (vingtile) bins of the distribution for third, fourth, and fifth grade 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 2002-03. 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 in the Classroom

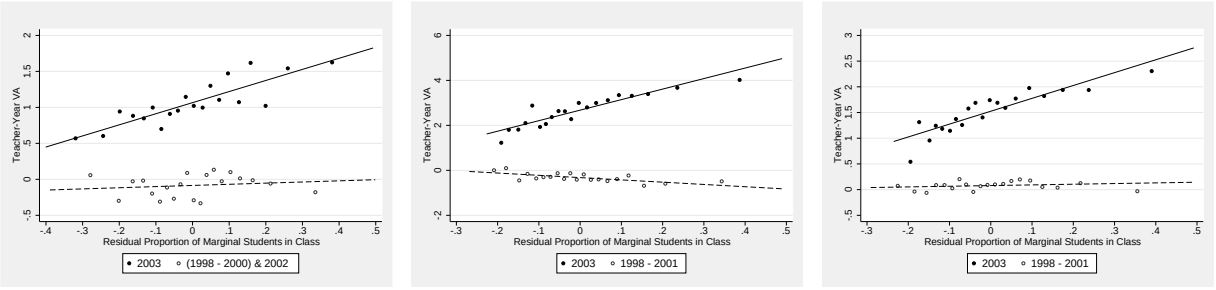


(a) All Grades (b) Third Grade

(c) Fourth Grade (d) Fifth Grade

Notes: The panels in this figure show the distributions of teachers' incentive-invariant abilities (which include baseline effort). To construct the figures, we estimate equation (6), 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.

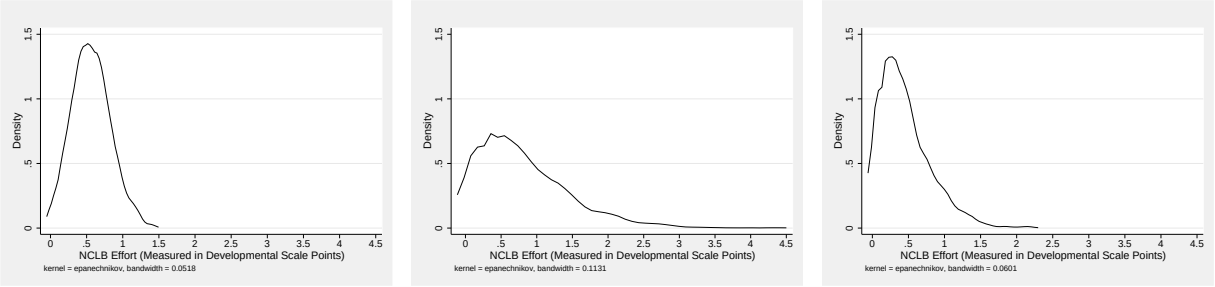
Figure 3: Incentive-Invariant Ability Distributions



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

Notes: This figure plots teachers' 2002-03 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 (7). For the pre-NCLB years, these controls also include year fixed effects. Accordingly, 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. For notational convenience, the legend in the panels labels profiles according to the latter year of the academic year in question. For example, the label '2003' identifies the profile corresponding to the 2002-03 academic year.

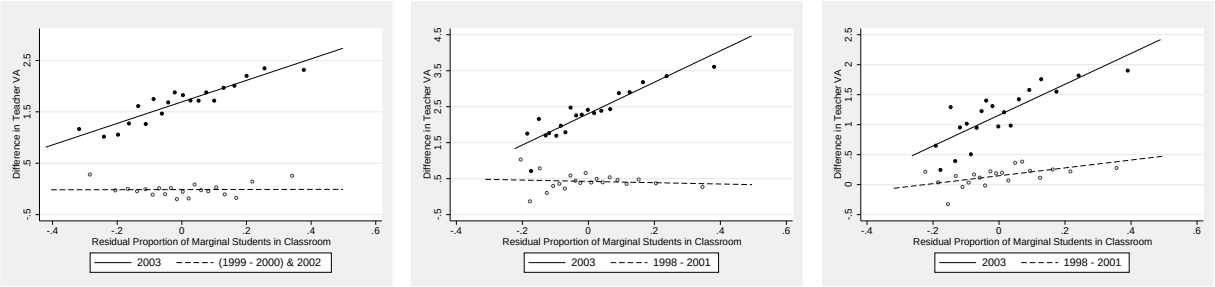
Figure 4: Teacher-Year VA versus Proportion of Marginal Students in 2002-03 Classrooms



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

Notes: This figure illustrates teachers' 2003 effort responses. In panels (a) to (c), we present grade-specific densities of 2002-03 effort levels. To construct these figures, we first obtain 2002-03 effort for each teacher by taking the linear prediction (fitted value) from $e(m_{j2003}) = \hat{\psi} m_{j2003}$. We then plot the distributions of these effort levels separately by grade.

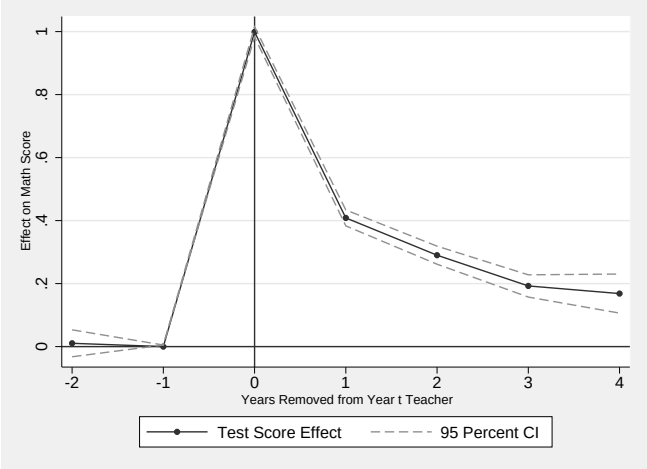
Figure 5: Effort Distributions in 2002-03



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

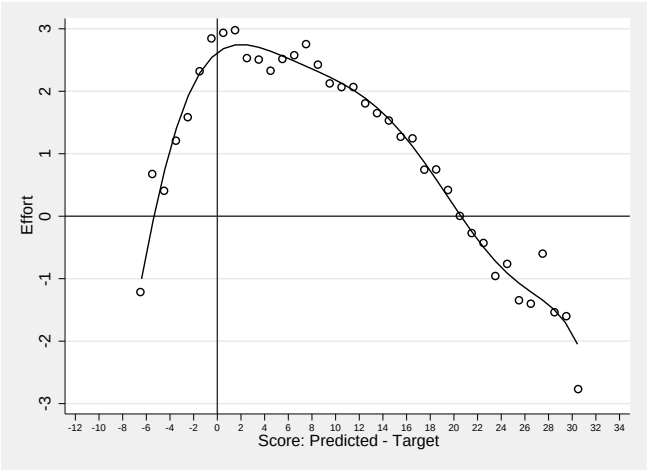
Notes: This figure illustrates teachers' 2002-03 effort responses. In panels (a) to (c), we depict the partial relationship between the change in teachers' annual performance from 2001-02 to 2002-03 and the fraction of students in their classes who were marginal in 2002-03. We also depict the partial relationship between the change in teachers' annual performance from all years $t-1$ to t in the pre-NCLB period and the fraction of students in their classes who were marginal in year t. To construct the panels, we first residualize m_{jt} with respect to the other controls in equation (9). For the pre-NCLB years, these controls also include year fixed effects. Accordingly, 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 change from years $t-1$ to t between each teacher's teacher-year fixed effects. The circles in each panel represent these averages. The lines represent the associated linear effects, estimated on the underlying teacher-year data. For notational convenience, the legend in the panels labels profiles according to the latter year of the academic year in question. For example, the label '2003' identifies the profile corresponding to the 2002-03 academic year.

Figure 6: Within-Teacher Performance Improvements



Notes: This figure reports estimates of the ϕ_n coefficients from equation (10). Each estimate is obtained from a separate regression in the pre-NCLB period from 1996-97 to 2001-02. The horizontal axis measures the number of years separating students 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



Notes: The horizontal axis is the same as in Figure 1. To construct this figure, we first take the bin-specific differences between the year 2002-03 and the year 1999-00 vertical-axis variable in Figure 1. The dots represent the resulting within-bin differences. We then estimate an eight-order polynomial using the binned data, weighting the regression by the number of student observations (across both 2002-03 and 1999-00) within each bin.

Figure 8: Student-Specific Effort Function

Appendices

Appendix A Data: Implications of Test Score Scale Changes

In this appendix, we discuss the timing of the changes to the developmental scale that the mathematics and reading end-of-grade tests are measured on. We also draw out the implications of these changes for our analysis, particularly regarding our methods for estimating teacher VA, predicting student test scores, and estimating the contemporaneous and persistent effects of teacher effort.[43]

A.I Mathematics Scale Change

Mathematics scores are measured on different scales before and after 2000-01. Because North Carolina's ABCs accountability program requires test scores from adjacent years in order to calculate student growth, the state provided a conversion table for the test scales. We convert 'second edition' scale scores to their 'first edition' counterparts for all tests except the third grade pre-test, which is written at the start of the academic year. The state did not provide a conversion table for the third grade pre-test because both the pre-test and the end-of-grade test would be on the second edition scale in 2000-01, thus making it possible to calculate student growth. We express all level and gain mathematics score summary statistics on the first edition scale, except third grade gains, which are calculated using first edition scores prior to 2000-01 and second edition scores for 2000-01 onwards.

A.I.i Implications of Scale Change for Estimating Teacher Value-Added

Teacher-Year Fixed Effects: When estimating teacher-year fixed effects for fourth and fifth grade, we convert all post 2000-01 test scores back to the first edition scale and estimate teacher VA using a pooled regression covering 1996-97 to 2004-05 (equation (B.1) in Appendix B below). Here, both contemporaneous test scores (the dependent variable) and lagged test scores (control variables) are on the first edition scale, allowing us to measure fourth and fifth grade teacher VA on the first edition scale throughout the sample period. For third grade, as mentioned above, we are not able to convert the second grade test to the first edition scale. Because the prior-year test score is needed as a control variable in the value-added estimation, we conduct two separate regressions: prior to 2000-01, third grade teacher value-added is measured on the first edition scale, while post 2000-01, it is measured on the second edition scale.

[43] Full discussions of these topics appear in Section V and Appendix B (for the estimation of teacher VA) and in Section VI and Appendix D (for predicting student test scores and for estimating the effects of effort).

Empirical Bayes Estimates of Teacher Ability: When estimating incentive-invariant teacher ability using the Empirical Bayes (EB) procedure in the pre-NCLB period (equation (6) in the main text), the differential timing of the 2000-01 mathematics developmental scale change in third grade and the non-availability of second grade scores in 1995-1996 together imply that we have two fewer years of data for third grade teachers than for fourth and fifth grade teachers. We therefore estimate a separate EB regression for third grade teachers, where the sample period runs from 1997-98 to 1999-00 (instead of from 1996-97 to 2000-01, as for the pooled regression of students and teachers in fourth and fifth grade).

A.I.ii Implications of Scale Change for Test Score Prediction

As explained in Section VI of the main text and Appendix D below, part of our empirical approach involves predicting student test scores in both the year NCLB was introduced – 2002-03 – and in years prior. We predict scores in prior years in order to conduct placebo tests when NCLB incentives were not operating. To estimate our prediction equations and categorize students according to their predicted scores, we opt *not* to use converted (across first and second edition scales) test scores. Instead, we measure test scores on the scale that was in effect when the tests were written.[44] This prevents us from using a prediction equation estimated prior to 2000-01 to predict test scores in 2000-01 and after. As a result, for fourth and fifth grades, the nearest pre-NCLB year for which we have predicted scores to conduct placebo tests is 1999-00. For third grade, we are able to conduct placebo tests using data from 2001-02. Here, we rely on the third grade pre-test (second grade test) and the end-of-grade test both being written and measured on the same scale in 2000-01. We then use data from 2000-01 to estimate the prediction equation, using that equation and out-of-sample student covariates in 2001-02 to predict performance in 2001-02.

A.I.iii Implications of Scale Change for Estimating the Effort Function and Effort Persistence

As discussed in Section VI of the main text and Appendix D below, we use the difference between students' realized and predicted scores in 2002-03 to estimate student-level effort when NCLB is introduced, and the difference between students' realized and 'counterfactual' predicted scores in 2003-04 to estimate the persistence of effort one year forward. Because we do not convert second edition scores back to the first edition scale when constructing predicted scores, we must use realized scores that are measured using the second edition when conducting these exercises. Therefore, both initial NCLB effort and its persistence are estimated using mathematics test scores measured on the second edition scale.

[44] Conversion across scales results in lumpiness in the distribution of predicted scores because the mapping from second to first edition scales is many-to-one – that is, in some instances, more than one test score on the second edition scale corresponds to the same value on the first edition scale. Because it is important for us to classify students correctly according to the distance between their predicted score and the proficiency cutoff, we conduct the analysis without converting.

A.II Reading Scale Change

We do not conduct our main analyses with reading scores because the scale used to measure reading tests changed in 2002-03, coinciding exactly with the introduction of NCLB. In addition, because we opt not to convert scores between first and second editions when categorizing students based on predicted scores, the timing of the scale change prevents us from identifying 'marginal' students based on reading scores in 2002-03.

Appendix B Estimating Teacher Value-Added: Technical Details

This appendix describes how we select the sample of students and teachers for teacher VA estimation, how we estimate teacher-year fixed effects, and how we interpret Empirical Bayes (EB) estimates of teacher ability in light of North Carolina's pre-existing ABCs program.

B.I Construction of the Teacher Value-Added Sample

To estimate teacher value-added (and to explore the separate effects of teacher ability and effort), we need to match students in the end-of-grade (EOG) files to their teachers in an accurate way in any given year. Using data on students and teachers from 1996-97 to 2004-05, we follow previous studies that use the NCERDC data by restricting attention to students in third through fifth grade, given that the teacher recorded as the test proctor is typically the teacher who taught the students throughout the year. We 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 taught a self-contained class for the relevant grade in the relevant year and if at least half of the tests administered by that teacher were 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.

When calculating value-added for each teacher, we include a given year of performance data in the value-added regressions for that teacher only if she had more than seven but fewer than forty students in her class with valid test scores and demographic variables, following the existing literature that has used the North Carolina data. A student is excluded from the value-added analysis if any of the following conditions hold: (1) the student had multiple scores for current or lagged EOG mathematics or readings tests; (2) the student had EOG scores corresponding to two or more teachers in a given year; (3) the student had EOG scores corresponding to two or more grades in a given year; or (4) the student had EOG scores corresponding to two or more schools in a given year. Applying these restrictions leaves 1.67 million student-year observations for estimating teacher VA. The summary statistics for this sample are presented in Table B.1 below.

B.II Estimation of Teacher-Year Fixed Effects

As described in Section V, we compute teacher-year fixed effects for each teacher using all students and teachers who are in the VA estimation sample. In doing so, we follow recent studies and regress contemporaneous mathematics scores on prior mathematics and reading scores and other student characteristics, in addition to the teacher-year fixed effects that are our main focus. The scores we use are measured on a developmental scale, rather than being standardized as is common in the literature (usually at the

grade-year level).[45] Here we rely on the careful psychometric design of developmental scales in a North Carolina context, which ensures that one can track improvements (or declines) in learning within and across students as they progress through school.

To estimate teacher-year fixed effects, we specify the following grade-specific regressions (for third, fourth and fifth grades):

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

Equation (B.1) is the empirical analog to equation (3) in the main text. In bringing the latter to the data, we control flexibly for the lagged test score, letting $f(y_{i,j',g-1,s',t-1})$ be a cubic function of lagged mathematics and reading scores; teacher-year fixed effects are denoted by q_{jt}, and we include a host of other determinants of test scores (abstracted from in the conceptual framework).[46] Conditional on those covariates, we obtain teacher-year fixed effect estimates as

$$\hat{q}_{jt} = \sum_{i=1}^{n(j,t)} \frac{y_{ijgst} - \hat{f}(y_{i,j',g-1,s',t-1}) - x'_{ijgst}\hat{\beta}}{n(j,t)}, \tag{B.2}$$

where $n(j,t)$ denotes the number of students in teacher j's classroom in academic year t. The resulting estimates represent a teacher's average contribution to her students' test scores, along with a common classroom shock that includes mean test score noise ($\bar{\epsilon}_{jt}$), thus providing the basis for equation (5) in Section V. Summary statistics for our estimated teacher-year fixed effects are presented in Table B.2 below.

B.III Interpreting the Estimates of Teacher Ability with the ABCs

In this section, we interpret our estimates of incentive-invariant ability in light of North Carolina's pre-existing ABCs program. In Section V, we used the EB estimator to recover teacher ability, estimating the following pooled regression across grades in the pre-NCLB period:

$$y_{ijgst} = f_g(y_{i,j',g-1,s',t}) + x'_{ijgst}\beta + h(exp_{jt}) + a_j + \theta_{jt} + \epsilon_{ijgst}, \tag{B.3}$$

where a_j represents teacher ability, θ_{jt} is a classroom-specific shock, and ϵ_{ijgst} is student-level noise.

Taking into account the fact that that the ABCs program was already operating in the pre-NCLB

[45] Although standardizing test scores guards against changes in testing regimes over time, de-meaning would effectively remove the effects of changes in performance incentives; and given our goal of assessing how teacher effort affects student learning, we wish to preserve all incentive-related performance variation over time.

[46] The other controls, x_{ijgst}, serve to mitigate the bias caused by non-random sorting of students to teachers (Chetty et al. 2014a). They consist of student race, gender, disability status, limited English-proficiency classification, parental education, and an indicator for grade repetition, and are likely to be correlated with innate student ability and previous teacher assignments.

period, we can estimate the same equation but modify the notation to

$$y_{ijgst} = f_g(y_{i,j',g-1,s',t}) + x'_{ijgst}\beta + h(exp_{jt}) + \mu_j + \theta_{jt} + \epsilon_{ijgst}, \tag{B.4}$$

where $\mu_j \equiv a_j + \underline{e_j}$ is the sum of true incentive-invariant teacher ability (a_j) and a term we label 'baseline' ABCs effort ($\underline{e_j}$). Baseline effort reflects the average ABCs-related effort exerted by the teacher across all of her years of teaching under the ABCs program. We cannot identify incentive-invariant ability and baseline effort separately, instead estimating a composite of the two, $\widehat{(a_j + \underline{e_j})}$, for each teacher j.

Our strategy for identifying the variation in teacher performance in 2002-03 that is driven by NCLB incentives relies on an across-teachers comparison and should not be affected by our inability to separate true incentive-invariant ability from baseline effort. This follows from the fact that the ABCs sets only a school-level target, without associated student-level test-score proficiency thresholds[47] – a design feature that contrasts sharply with NCLB and ensures that effort incentives under the ABCs operate at the school level.[48] As such, the baseline effort component of our EB estimate is likely to be constant across teachers within a school, implying that our estimates reflect true incentive-invariant ability plus a constant shift (to all teachers within a school). As there is little variation in ABCs effort incentives across teachers, the pre-existing accountability program does not confound our estimation of the effects of NCLB incentives on teacher performance, given that our identification strategy exploits variation in NCLB incentives *across* teachers in 2002-03.

It is unlikely that the introduction of NCLB in 2002-03 created systematic variation in ABCs incentives across teachers, as the rules of the ABCs remained constant in that year. Our identification strategy does permit aggregate changes to ABCs incentives at the school level – for example, by NCLB drawing attention away from ABCs-related considerations. To separately identify the variation in performance due to NCLB incentives from the variation due to ABCs incentives, we only require that ABCs-related incentives do not change across teachers in a way that is correlated with the strength of NCLB incentives in 2002-03.

[47] The ABCs sets average growth targets at the grade level and then aggregates the differences between average and target growth across all grades within a school to arrive at a school-level growth score. Under the ABCs, average test score growth at a school is the key determinant of school success under the program, regardless of where the growth is concentrated in terms of the underlying student distribution.

[48] For a more detailed discussion of the ABCs, see Macartney (2016).

Table B.1: Student-Level Summary Statistics for Value-Added Sample

	Mean	SD	Observations
Performance Measures			
Mathematics Score			
Grade 3	145.09	10.49	595,097
Grade 4	154.10	9.56	553,833
Grade 5	160.48	9.15	527,762
Mathematics Growth			
Grade 3	13.90	6.30	595,097
Grade 4	9.34	6.02	553,833
Grade 5	7.13	5.28	527,762
Future[(a)] Mathematics Score			
Grade 6	167.84	10.76	456,348
Grade 7	173.29	10.41	387,525
Grade 8	176.43	11.07	316,557
Reading Score			
Grade 3	147.37	9.19	595,097
Grade 4	150.95	9.02	553,833
Grade 5	156.21	7.93	527,762
Reading Growth			
Grade 3	8.24	6.70	595,097
Grade 4	3.90	5.55	553,833
Grade 5	5.56	5.20	527,762
Future[(a)] Reading Score			
Grade 6	157.60	8.36	455,871
Grade 7	161.29	7.63	387,140
Grade 8	163.84	7.16	316,225
Demographics			
College-Educated Parents	0.26	0.44	1,676,692
Male	0.50	0.50	1,676,692
Minority	0.37	0.48	1,676,692
Disabled	0.05	0.22	1,676,692
Limited English-Proficient	0.02	0.13	1,676,692
Repeating Grade	0.01	0.10	1,676,692
Free or Reduced-Price Lunch	0.41	0.50	1,203,519

Notes: Summary statistics are calculated for all third through fifth grade student-year observations from 1996-97 to 2004-05.
[(a)] Future mathematics and reading scores are the scores we observe for our sample of third through fifth grade students when they are in sixth, seventh, and eight grades. 'Future' mathematics and reading scores are used when measuring the persistent effects of teacher ability and effort. We do not follow students past 2004-05, as the mathematics scale changes again in 2005-2006 yet no table to convert scores back to the old scale was created by the state. The free or reduced-price lunch eligibility variable is not available prior to 1998-99.

Table B.2: Teacher-Year Fixed Effects Summary

	(1)	(2)	(3)
Grade	3rd	4th	5th
Mean	-0.17	-0.11	0.19
Standard Deviation	2.65	2.80	2.31
Observations	24,105	22,246	20,596

Notes: This table presents means and standard deviations for the teacher-year fixed estimates. Summary statistics are calculated using all available teacher-year observations from 1996-97 to 2002-03.

Appendix C Robustness Checks

In this Appendix, we show that the results in Section V.B are robust to alternative ways of defining students as marginal. We also rule out two plausible rival hypotheses to teacher effort setting as driving our main results – namely, differential sorting of students to teachers by ability and differential class size adjustments in response to NCLB.

C.I Alternative Definitions of 'Marginal Student'

We first demonstrate that the patterns in Figure 4 are robust to alternative cut-offs for defining a student as marginal. Figure C.1 below shows teacher VA in each grade as a function of the fraction of marginal students in the classroom in 2002-03 and a placebo year (for many different definitions of 'marginal students'). Each panel of Figure C.1 shows an increasing relationship in 2002-03 and no relationship in the placebo year, lending credence to our claim that our results do not depend on the way we choose to classify students as marginal.

C.II Rival Hypotheses to Teacher Effort Setting

Our leading hypothesis is that the measured test score improvement is due to an increase in teacher effort in response to the incentives under the proficiency count system. Given that effort is not observed directly, it is important to consider whether the evidence might be consistent with alternative hypotheses. In Section V.B.3, we summarize two such hypotheses, which are explained in greater detail here.

Differential Sorting by Ability: One rival explanation is that students were sorted *differentially* to teachers in 2002-03, with high (incentive-invariant) ability teachers receiving greater fractions of marginal students. While we do 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 a relationship between teacher and student ability could be driving the results rather than additional effort being exerted by a *given* teacher.

A natural way to assess this rival hypothesis is to test whether the relationship between the fraction of marginal students in a classroom and teacher ability changes in 2002-03. We conduct this test by regressing the fraction of marginal students in each class on grade and year fixed effects (λ_g and λ_t, respectively), our measure of teacher incentive-invariant ability, and an interaction of that term with a 2002-03 indicator:

$$m_{jt} = \alpha_0 + \lambda_g + \lambda_t + \beta_1 \hat{a_j} + \beta_2 \hat{a_j} \times 1(t = \text{2002-03}) + \epsilon_{jt}, \ \forall \ t \leq \text{2002-03}. \tag{C.1}$$

If principals began sorting students to teachers differentially on the basis of ability in 2002-03, we would expect to find a non-zero β_2 coefficient.

Table C.1 shows the results from estimating variants of equation (C.1). Overall, there is a small *negative* relationship between the fraction of marginal students who are in a teacher's class and the teacher's incentive-invariant ability. This reflects the relatively low test score proficiency standard in North Carolina and the sorting of low-performing students to low-ability teachers.[49] The sorting pattern appears to change slightly in 2002-03, but indicates that high-ability teachers received *smaller* fractions of marginal students than in the pre-NCLB period.[50] For our main results to be biased upward, the change would need to be in the opposite direction.

The low test score proficiency target and the sorting patterns of students to teachers together result in slightly stronger effort incentives for low-ability teachers. Figure C.2 shows grade-specific relationships between $e(m_{jt})$ and a_j. For each grade, we plot the relationships that prevail in 2002-03 and the pooled pre-NCLB control years. In 2002-03, there is a clear *decreasing* relationship between NCLB effort and incentive-invariant ability: in the control years, the estimated functions are virtually flat. The slope coefficients in 2002-03 are -0.01, -0.04, and -0.02, for third, fourth and fifth grade, respectively, and are all significant at the one percent level. Thus, a one standard deviation *lower* ability teacher in 2002-03 exerted approximately 0.02, 0.07, and 0.03 developmental scale points worth of additional effort in third, fourth grade, and fifth grades. These differences are very small, however, corresponding to 0.002, 0.007, and 0.003 student-level standard deviations.

Differential Sorting by Class Size: We also assess the robustness of our results to the possibility that schools might sort marginal students differentially into smaller sized classrooms in response to NCLB. Such a response could arise if schools thought that marginal students might perform better there. In Table C.2, we investigate the importance of class size by including it as a control variable and replicating the analysis in panel (a) of Table 3. None of the point estimates in the table are statistically distinguishable from their Table 3 counterparts.

In sum, the evidence argues against the hypothesis that it is differential sorting – either by ability or class size – that is driving our main results. We view this as clear support for our effort interpretation.

[49]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).

[50]To put the magnitude of the change in perspective, a teacher who is one standard deviation (1.79 developmental scale points) below average had (1.79*0.0033) 0.59 percentage points more marginal students in her classroom in the post-NCLB period. This corresponds to 2 percent of the classroom-level mean fraction of marginal students.

Table C.1: Tests for Differential Sorting of Students to Teachers in 2002-03

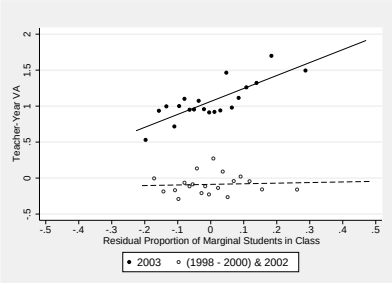
	(1) Full Sample	(2) Third Grade	(3) Fourth Grade	(4) Fifth Grade
Ability	-0.0034***	-0.0010	-0.0046***	-0.0055***
	(0.0008)	(0.0010)	(0.0011)	(0.0017)
$1(t = 2003) \times$Ability	-0.0033**	-0.0050***	-0.0045**	0.0027
	(0.0014)	(0.0019)	(0.0019)	(0.0025)
Observations	39,932	12,599	14,151	13,182

Notes: This table presents the results of regressions based on equation (C.1). 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 (6), 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.

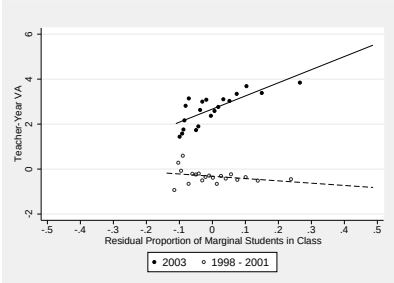
Table C.2: The Effects of NCLB Incentives Controlling for Class Size

Dependent Variable: Teacher-Year VA						
	Third Grade		Fourth Grade		Fifth Grade	
	2002-03	Pre-NCLB	2002-03	Pre-NCLB	2002-03	Pre-NCLB
Effect of m_{jt}	1.39***	-0.02	4.16***	-0.99***	2.25***	-0.13
	(0.21)	(0.15)	(0.32)	(0.15)	(0.24)	(0.15)
Observations	2,144	10,452	2,598	11,551	2,570	10,609

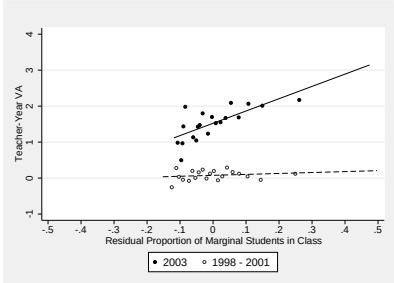
Notes: This table presents estimates of ψ from grade-specific regressions of equation (7). In the year 2002-03 regression, additional controls include teacher ability, teacher experience, and class size. The result in the pre-NCLB columns comes from a pooled regression of all pre-NCLB years that additionally 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. 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.



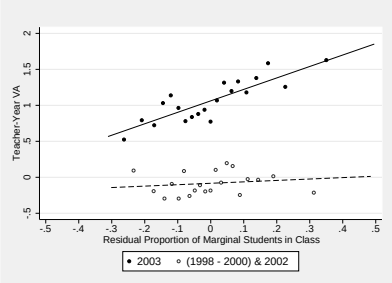
(a) 3rd Grade: $-2 \leq \hat{y} - y^T \leq 2$



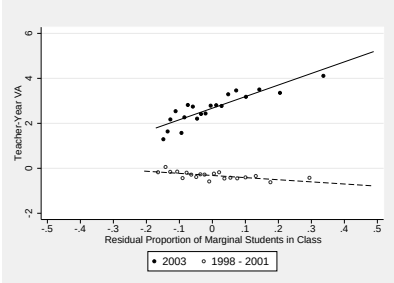
(b) 4th Grade: $-2 \leq \hat{y} - y^T \leq 2$



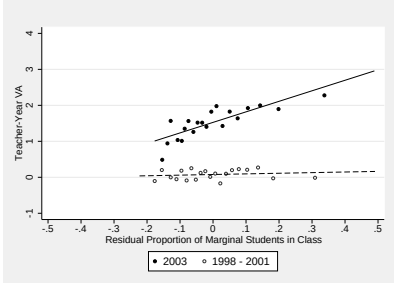
(c) 5th Grade: $-2 \leq \hat{y} - y^T \leq 2$



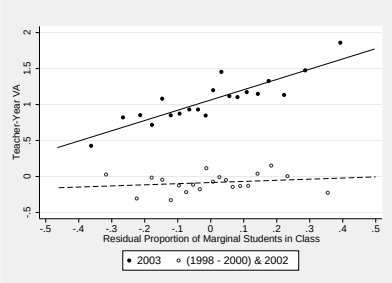
(d) 3rd Grade: $-3 \leq \hat{y} - y^T \leq 3$



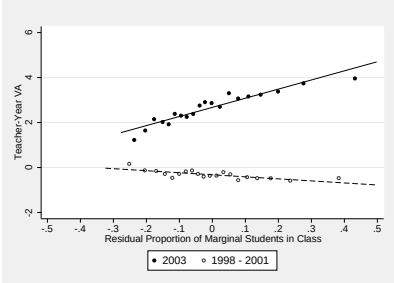
(e) 4th Grade: $-3 \leq \hat{y} - y^T \leq 3$



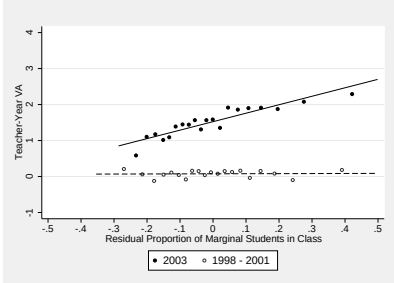
(f) 5th Grade: $-3 \leq \hat{y} - y^T \leq 3$



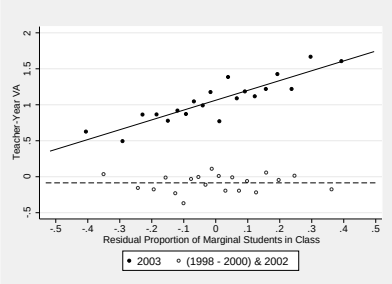
(g) 3rd Grade: $-5 \leq \hat{y} - y^T \leq 5$



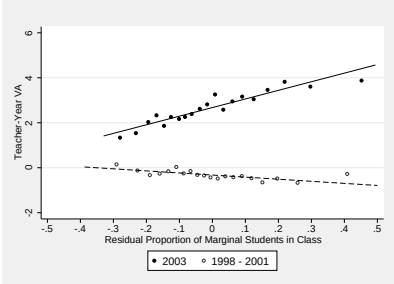
(h) 4th Grade: $-5 \leq \hat{y} - y^T \leq 5$



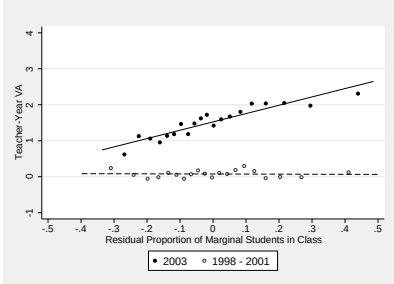
(i) 5th Grade: $-5 \leq \hat{y} - y^T \leq 5$



(j) 3rd Grade: $-6 \leq \hat{y} - y^T \leq 6$



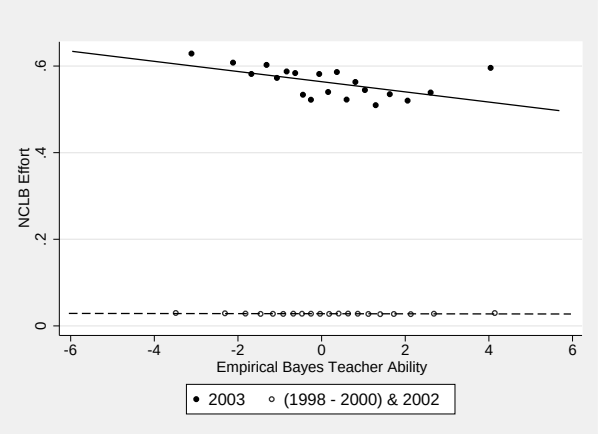
(k) 4th Grade: $-6 \leq \hat{y} - y^T \leq 6$



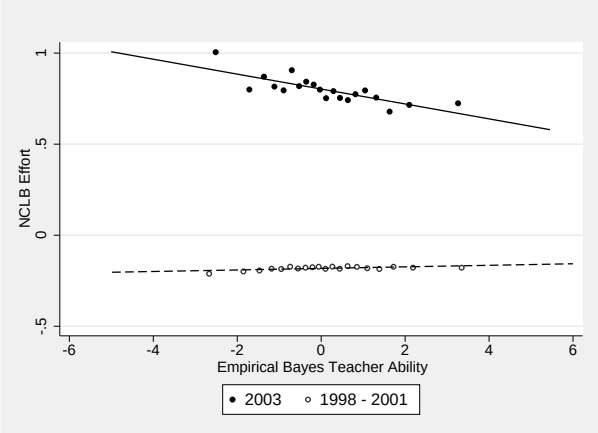
(l) 5th Grade: $-6 \leq \hat{y} - y^T \leq 6$

Notes: This figure reproduces the analysis in Figure 4 with alternative cutoffs for marginal student status. See the notes of Figure 4 for details. The range for marginal student classification is given in the label of each panel.

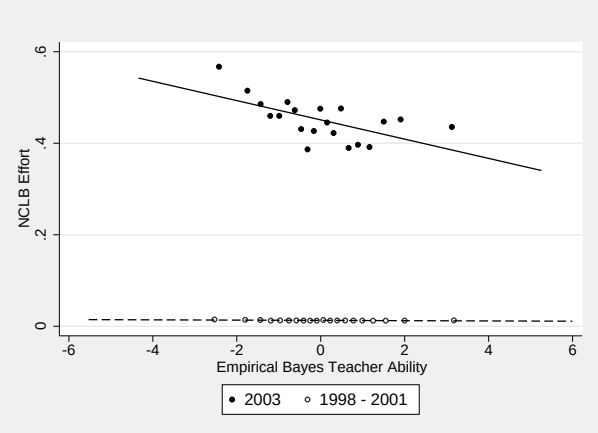
Figure C.1: Effort Predictions in 2002-03 with Alternative Definitions of 'Marginal Student'



(a) Third Grade



(b) Fourth Grade



(c) Fifth Grade

Notes: This figure plots the relationship between NCLB teacher effort, obtained as the fitted value from from $e(m_{jt}) = \hat{\psi} m_{jt}$, 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 C.2: The Relationship Between Teacher Effort and Incentive-Invariant Ability

Appendix D Estimating the Persistence of Effort: Technical Details

This appendix presents a detailed discussion of the methodology we outline in Section VI.B for estimating the persistence of teacher effort. We first present a summary of the key variables used in our approach. The notation in hand, we then describe how we form an estimate of teacher effort at the student level (Section VI.B.1 in the main text) and how we derive the main estimating equation (Section VI.B.2). We also describe our estimation strategy (Section VI.B.3).

D.I Definitions

Table D.1 below presents the definitions of key variables used to exposit our approach for estimating the persistence of effort.

Table D.1: Estimating the Persistence of Effort: Relevant Notation

Variable	Definition
$\hat{y}_{i,j,g,s,02-03}$	The predicted mathematics score of student i who is assigned to teacher j in grade g at school s in academic year 2002-03. This predicted score is calculated according to the steps outlined in Section D.II below.
$y_g^{T,N}$	The mathematics test score proficiency target in grade g mandated by NCLB. The superscripts T and N indicated 'target' and 'NCLB,' respectively.
$\pi_{i,02-03}$ $(\equiv \hat{y}_{i,j,g,s,02-03} - y_g^{T,N})$	Incentive strength in 2002-03: The difference between the predicted mathematics score of student i and the test score proficiency target mandated by NCLB in 2002-03.
$y^C_{i,j,g,s,03-04}$	The 'counterfactual' predicted mathematics score of student i who is assigned to teacher j in grade g at school s in academic year 2003-04. This predicted score is calculated according to the steps outlined in Section D.III below.
$\pi_{i,03-04}$ $(\equiv y^C_{i,j,g,s,03-04} + \gamma_1^e e^N(\pi_{i,02-03}) - y_g^{T,N})$	Incentive strength in 2003-04: The difference between the predicted mathematics score of student i and the test score proficiency target mandated by NCLB in 2003-04.
$e^N(\cdot)$	The empirical effort function shown in Figure 8. In academic year 2002-03, this function takes $\pi_{i,02-03}$ as its argument. In academic year 2003-04, it takes $\pi_{i,03-04}$ as its argument.
γ_1^e	The one-period forward persistence rate of NCLB effort.

D.II Estimating Student-Level Effort

As noted in Section VI.B.1, we begin by constructing a student-level measure of effort, which draws on the non-parametric patterns in Figure 1. The figure shows that the introduction of NCLB had clear non-linear effects on student test scores, consistent with strong teacher effort responses to the scheme. This interpretation of Figure 1 depends on two key concepts, namely students' predicted scores and a measure of incentive strength. We discuss each in turn, followed by an explanation of how they are used together to construct student-level effort.

D.II.i Predicted Scores

The predicted score for each student in 2002-03 captures the score students would earn in that year had NCLB not been enacted. Empirically, we calculate the predicted score in two steps:

Step 1 – Estimate a Prediction Equation in the Pre-NCLB Period: We predict student performance in a flexible way in pre-NCLB years using several covariates. Specifically, we regress contemporaneous 2001-02 mathematics scores on cubics in prior 2000-01 mathematics and reading scores and indicators for parental education, gender, race, free or reduced-price lunch eligibility, and limited English proficiency. We then save the coefficients from this regression.

Step 2 – Use the Prediction Equation with Student Covariates in 2002-03: We make an out-of-sample mathematics test score prediction for students in 2002-03, denoting the predicted score for student i in 2002-03 by $\hat{y}_{i,j,g,s,02-03}$. We calculate it by combining the estimated coefficients from the first step with the (pre-determined) covariates of students in 2002-03.

Because it is estimated using the relationship between student characteristics and test scores that prevailed prior to NCLB, the predicted score represents the test score students would have earned had teachers not adjusted their effort decisions in response to NCLB's introduction. Drawing on our conceptual framework in Section IV, we take the predicted score for student i to represent the score teachers predict student i would earn in 2002-03 without any additional NCLB-related effort. This prediction is based on all non-effort inputs, which we define formally using equation (3), setting effort to zero and assuming a mean-zero error:

Definition 1 – Predicted Student Score in 2002-03: $\hat{y}_{i,j,g,s,02-03} \equiv \gamma y_{i,j',g-1,s',01-02} + a_{j(i,02-03)}$.

With this definition in hand, we use the predicted score together with students' realized test scores in 2002-03 ($y_{i,j,g,s,02-03}$) to obtain a (noisy) estimate of the effort received by each student. In particular,

take the difference between the realized and predicted score for a given student:

$$\begin{aligned} y_{i,j,g,s,02-03} - \hat{y}_{i,j,g,s,02-03} &= \gamma y_{i,j',g-1,s'01-02} + a_{j(i,02-03)} + e_{j(i,02-03)} + \epsilon_{i,j,g,s,02-03} \\ &\quad - (\gamma y_{i,j',g-1,s',01-02} + a_{j(i,02-03)}) \\ &= e_{j(i,02-03)} + \epsilon_{i,j,g,s,02-03}. \end{aligned} \tag{D.1}$$

This results in the sum of NCLB effort and a random shock to test scores. As we describe below, in order to eliminate the influence of test score noise, the vertical axis in Figure 1 plots averages of the differences given by equation (D.1).

D.II.ii Incentive Strength

The second quantity relevant for interpreting the patterns in Figure 1 is our measure of incentive strength, which depends on the distance between the predicted score and the fixed NCLB proficiency target. Letting $y_g^{T,N}$ (where 'N' in the superscript stands for 'NCLB') denote the NCLB test score proficiency target in grade g, we define incentive strength formally as the difference between the predicted score and the target:

Definition 2 – Incentive Strength in 2002-03: $\pi_{i,02-03} \equiv \hat{y}_{i,j,g,s,02-03} - y_g^{T,N}$.

This incentive strength measure is used on the horizontal axis in Figure 1, in which students are grouped into two-scale-point width bins of incentive strength in 2002-03. We then plot the *average* difference between the realized and predicted score (given by equation (D.1)) within each bin, eliminating idiosyncratic test score noise to recover average teacher effort as a function of incentive strength.

D.II.iii Using the Predicted Score and Incentive Strength to Estimate Student-Level Effort

Figure 1 makes clear that students predicted to score near the proficiency threshold ($\pi_{i,02-03} \approx 0$) – namely those for whom effort incentives are strongest – do receive the biggest boost to their scores. To ensure that we do not systematically under- or over-predict test scores for certain parts of the distribution, we conduct the same exercise in the 1999-2000 pre-reform period (when there is necessarily no NCLB effort response), showing that our predicted score tracks the realized score very well throughout the distribution, given by the flat line. This lends credence to the view that the 2002-03 patterns reflect student-specific NCLB effort.

We then use the profiles for the two years in Figure 1 to estimate a student-specific effort function that takes incentive strength as its argument. We do so by differencing the binned 2002-03 and 1999-00 profiles and then fitting an eighth-order polynomial to the differenced data using a weighted regression,

with the weights capturing the total number of students in each bin (across both 2002-03 and 1999-00). The resulting effort function, denoted by $e^N(\cdot)$ and constructed using the estimated coefficients from this regression, is plotted in Figure 8. We use this function to assign a level of effort to each student, according to the following assumption:

Assumption 3: Given student-specific values of $\pi_{i,02-03}$ and the function $e^N(\cdot)$, the effort directed to each student i in 2002-03 is given by $e_{j(i,02-03)} = e^N(\pi_{i,02-03})$.

D.III The Components of the Estimating Equation

With the student-specific effort measure in hand, we turn to specifying an equation that can be taken to the data for estimating the rate at which effort persists. As described in Section VI.B.2 of the main text, this equation depends on four components: (i) the counterfactual predicted score, (ii) student-level NCLB effort in 2002-03, (iii) student-level NCLB effort in 2003-04, and (iv) average school-level NCLB effort in 2002-03 (which captures the indirect effect of ABCs incentives on student test scores). While effort in 2002-03 is described above, we discuss the remaining three components in this section.

D.III.i The Counterfactual Predicted Score

Drawing on the technology presented in the conceptual framework, equation (4) allows us to specify how effort persists in influencing test scores one period ahead. Thus, adapting that equation, test scores in 2003-04 are determined by:

$$\begin{aligned} y_{i,j,g,s,03-04} = \gamma(y_{i,j',g-1,s,02-03} - a_{j(i,02-03)} - e_{j(i,02-03)}) \\ + a_{j(i,03-04)} + e_{j(i,03-04)} + \gamma_1^a a_{j(i,02-03)} + \gamma_1^e e_{j(i,02-03)} + \eta_{i,j,g,s,03-04}\,. \end{aligned} \tag{D.2}$$

The RHS of this equation captures (in turn) the persistent effect of once-lagged scores from 2002-03 excluding teacher ability or effort – the first term, $(y_{i,j',g-1,02-03} - a_{j(i,02-03)} - e_{j(i,02-03)})$ – the effects of teacher ability and teacher effort in the current year 2003-04, the persistent effects of teacher ability and effort from 2002-03, and a random shock to current test scores.

Under NCLB, marginal students are likely to receive the most teacher effort, as the descriptive evidence indicates. To identify marginal students in 2003-04, it helps to draw a distinction between the component of the predicted score in that year that is independent of the persistence of effort, which we call the 'counterfactual predicted score,' and the part that depends on it. We define the former as

Definition 3 – Counterfactual Predicted Score: $y^C_{i,j,g,s,03-04} \equiv \gamma(y_{i,j',g-1,s',02-03} - a_{j(i,02-03)} - e_{j(i,02-03)}) + a_{j(i,03-04)} + \gamma_1^a a_{j(i,02-03)}$.

This captures the test score that students *would* have earned in 2003-04 had NCLB not been enacted in the prior year, in which case there would be no contemporaneous or persistent effect of effort.

We construct an empirical analogue to the counterfactual predicted score by implementing a slight modification to our procedure outlined above for constructing the predicted score ($\hat{y}_{i,j,g,s,02-03}$). Specifically, for each grade, we estimated (prior to NCLB) the linear regression coefficients of a test-score prediction equation in Step 1 of the prediction exercise in Section D.II above. Step 2 of that exercise involves using those prediction equations (one for each grade) to make out-of-sample forecasts in 2002-03. Here, we use the equations to make forecasts one more year ahead, to the 2003-04 academic year, by substituting *predicted* test scores from 2002-03 ($\hat{y}_{i,j',g-1,s',02-03}$) in place of realized grade $g-1$ (i.e., prior-year) test scores. We use the actual 2003-04 values for all other covariates when making the forecast. Realized grade $g-1$ scores from 2002-03 contain NCLB effort, and their persistence into 2003-04 therefore depends on the parameter governing the persistence of effort, γ_1^e. Using predicted scores in place of realized scores ensures that the counterfactual predicted score represents the score students would earn in 2003-04 had there been no NCLB incentives in 2002-03.

D.III.ii Student-Level Effort in 2003-04

In order to obtain a measure of student-level NCLB effort in 2003-04, we must capture the way teachers form predictions about likely student performance in 2003-04, and, specifically, how they incorporate the persistence of prior-year effort into those predictions. To that end, we make the following pair of assumptions – about teachers' information sets and the test score prediction rule they follow, respectively:

Assumption 4a: Teachers know the level of effort devoted to each student in the previous year, $e_{j(i,02-03)} = e^N(\pi_{i,02-03})$, and the persistence rate of effort, γ_1^e.

Assumption 4b: The prediction teachers make about each student's contemporaneous test score (in 2003-04) in the absence of any contemporaneous effort is given by $y^C_{i,j,g,s,03-04} + \gamma_1^e e_{j(i,02-03)}$.

This prediction draws on the technology directly. To see this, according to equation (D.2) above, student test scores in 2003-04 are determined by the predicted score in that year plus contemporaneous teacher effort. The prediction is the sum of the counterfactual predicted score (which subsumes the contemporaneous and persistent effects of teacher ability and the portion of the prior score that is independent of teacher ability and effort) and the persistence of effort from the prior year. Teachers use these student-level predictions in 2003-04 when making decisions about how much effort to devote to each student, taking account of the incentive to direct effort to student i. That incentive is given in the next definition:

Definition 4 – Incentive Strength in 2003-04: $\pi_{i,03-04} \equiv y^C_{i,j,g,s,03-04} + \gamma_1^e e_{j(i,02-03)} - y_g^{T,N}$.

Specifically, teachers account for the distance, given by $\pi_{i,03-04}$, between a student's predicted score and the NCLB proficiency target, setting effort according to a rule captured in the following assumption:

Assumption 5: The effort devoted to student i in 2003-04 is given by $\theta e^N(\cdot)$ evaluated at $\pi_{i,03-04}$, where $\theta > 0$.

This 'shape' assumption implies that teachers use the same empirically-determined effort function as in 2002-03 to set effort, with the given function taking $\pi_{i,03-04}$ as its argument in 2003-04, and the parameter θ either diminishing (when $\theta < 1$) or amplifying (when $\theta > 1$) all effort levels in a proportional way. To justify this, it is plausible to think that teachers would direct effort to students in a similar way across the two years, with marginal students receiving relatively more effort than non-marginal students in each year, given that the rules of NCLB remained constant across 2002-03 and 2003-04.[51] The overall effort response across the two years might still differ if, for example, the novelty and added publicity of NCLB in its first year caused schools to try harder than they would in future years, with θ capturing changes in aggregate effort over time.[52]

Given Definition 4 and Assumption 5, it is clear that the effort decision in 2003-04 depends on the effort students received in 2002-03 and the effort persistence parameter γ_1^e, as these influence incentive strength in 2002-03, thus highlighting the correlation of effort over time. To obtain the estimating equation we take to the data, we also need to control for the way effort incentives under North Carolina's pre-existing ABCs program in 2003-04 were affected by effort responses following NCLB's introduction in 2002-03. We describe our approach for doing so next.

D.III.iii Accounting for ABCs Effort with Average School-Level Effort in 2002-03

Teacher effort responses to NCLB in 2002-03 can disrupt ABCs incentives because ABCs growth targets depend, by institutional design, on students' prior scores and do not discriminate between the (potentially) differential persistence rates of the inputs that contribute to those scores. Therefore, if effort persists at a *lower* rate than the ABCs target growth rate, then the target will grow faster than the test score, implying that the school-level ABCs target becomes more difficult to satisfy, in turn altering contemporaneous effort incentives. We use our conceptual framework to expand on this reasoning below and outline our strategy

[51] This assumption serves as an approximation to a more explicit modeling of the effort-setting process, as in the structural model in Macartney *et al.* (2015).

[52] As an alternative to having effort levels changing proportionally across years, one might think that over time, teachers would become better able to predict which students were marginal and thereby direct more effort to those students, resulting in a compressed effort function rather than one that is scaled up or down by a constant factor. This alternative hypothesis is unlikely in North Carolina, given that the state's pre-existing accountability program (the ABCs) relied on the same End-of-Grade tests and proficiency thresholds as NCLB; as the ABCs was implemented in 1996-97, educators had fully six years prior to NCLB to become familiar with the state tests and learn how to form expectations about student performance there. Teachers becoming better at predicting student proximity to the passing threshold is more likely to occur in states that did not have pre-existing accountability programs prior to NCLB.

for accounting for changing ABCs effort decisions.

North Carolina's ABCs program sets test score growth targets that are grade- and subject-specific for each school, and which are then aggregated across all grade-subject pairs within the school to form a school-level growth score. In the ABCs legislation, targets for average test score growth across all students in a subject-grade are set as a linear function of students' prior scores. A school passes the ABCs when the sum of the differences between average and target scores in each grade is greater than zero, the formal condition being written

$$\sum_{g=3}^{G_s} \sum_{i \in g} \frac{y_{ijgst} - \alpha y_{i,j',g-1,s',t-1}}{N_{gt}} \geq 0, \tag{D.3}$$

where G_s stands for the highest grade served at a given school, and α denotes the (pre-determined) coefficient that multiplies student prior scores to form test score growth targets.[53] In the equation, the first sum is taken over all grades in the school, from third grade up to grade G_s. The second sum is taken over all students in grade g in year t, and N_{gt} is the corresponding number of students in that grade-year combination in the school.

To see how an effort response to NCLB in 2002-03 can affect the likelihood of the school passing the ABCs in 2003-04, we use the test score technology and the notation established above to write the passing condition in 2003-04 as

$$\sum_{g=3}^{G_s} \left(\left(\bar{y}^C_{g,s,03-04} - \alpha \bar{\hat{y}}_{g-1,s,02-03} \right) + \bar{e}_{g,03-04} + (\gamma_1^e - \alpha) \bar{e}_{g-1,02-03} + \bar{\eta}_{g,s,03-04} - \alpha \bar{\epsilon}_{g-1,s,02-03} \right) \geq 0, \tag{D.4}$$

which consists of a sum of grade-specific averages.

As equation (D.4) shows, the extent to which the 2002-03 NCLB response disrupts 2003-04 ABCs incentives depends on how the persistence rate of effort compares with the target growth rate legislated under the ABCs, given by the difference, $(\gamma_1^e - \alpha)$. The target takes the 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 α, then the target grows at a faster rate than the tests score, implying that the school-level ABCs target is more difficult to satisfy than in the pre-NCLB period. The opposite is true when effort persists at a *higher* rate than α. In either case, schools may respond to the change in ABCs incentives by adjusting contemporaneous effort. In order to separately identify the direct persistent effect of once-lagged effort on student test scores, we need to account for such responses.

[53] The essential feature of the actual targets set under the ABCs is that they are *linear* in the prior scores. 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 mathematics and reading scores, and individual multiplicative coefficients for each.

We do so by recognizing that, for a given persistence rate of effort and ABCs required growth rate, *average* 2002-03 NCLB effort is the key determinant of the distortion to 2003-04 school-level ABCs incentives.[54] In particular, after moving the summation through (D.4), the distortion is captured by $(\gamma_1^e - \alpha)\sum_{g=3}^{G} \bar{e}_{g-1,02-03}$. When estimating the persistence of NCLB effort in 2002-03, we therefore control for average school-level 2002-03 NCLB effort to account for distortions to ABCs incentives. In particular, we make the following assumption:

Assumption 6: In 2003-04, the effect of the effort response to NCLB in 2002-03 on ABCs effort incentives (parameterized by ρ) is determined by the average school-level effort response from the prior year, $\bar{e}^N_{s,02-03}$, with the change to ABCs effort incentives being common to all students in a given school.

Because schools' ABCs effort responses in 2003-04 are unobserved, our strategy does not control for the responses directly; instead, we hold the effort response fixed by accounting for the variable that determines the effort response through its effect on the degree to which a school alters its probability of passing the ABCs.

D.III.iv The Estimating Equation

Using the notation we have now introduced, we use the test score technology described in equation (D.2) and the definition of the counterfactual predicted score ($y^C_{i,j,g,s,03-04}$) to write the main estimating equation as:

$$y_{i,j,g,s,03-04} - y^C_{i,j,g,s,03-04} = \gamma_1^e e^N(\pi_{i,02-03}) + \theta e^N(\pi_{i,03-04}) + \rho \bar{e}^N_{s,02-03} + \eta_{i,j,g,s,03-04}. \tag{D.5}$$

Deducting the counterfactual predicted score from students' realized scores on the LHS allows us to isolate all the NCLB effort responses that are relevant from an estimation perspective on the RHS. The first term – effort in 2002-03 – is known by the econometrician, determined by incentive strength, $\pi_{i,02-03}$, according to the semi-parametric effort function $e^N(\cdot)$. The second term, unknown to the econometrician, is effort in 2003-04, which depends on the key persistence parameter of interest. The third term, following Assumption 6, adds average school effort from 2002-03, multiplied by parameter ρ, to control for ABCs effort incentives in 2003-04.[55]

[54] There is no distortion only when effort persists at the same rate as the ABCs required growth rate, i.e. when $\alpha = \gamma_1^e$.

[55] We calculate $\bar{e}^N_{s,2003}$ as the jack-knife mean 2002-03 effort across all students in school s, leaving out the effort received by student i, to ensure that the estimates of γ_1^e and ρ are not confounded.

D.IV Estimation Approach and Identification

We now provide more detail about the estimation approach we outlined in Section VI.B.3 and further discuss identification of the main parameters of interest.

D.IV.i Estimation Approach

Our goal, having specified the estimating equation, is to recover effort persistence (γ_1^e), the scale factor on contemporaneous effort (θ), and the indirect effect of ABCs incentives on student test scores (ρ). The main challenge is that the input to the 2003-04 effort function depends on the (unknown) persistence rate. Yet in order to estimate the persistence rate, we need to account for the correlation of effort across time. Thus we seek to estimate effort received by students in 2003-04 *simultaneously* with the parameters of interest.

To that end, we use a maximum likelihood approach, making the following distributional assumption about the error in equation (D.5):

Assumption 7: $\eta_{i,j,g,s,03-04} \sim N(\mu, \sigma^2)$.

Equation (D.5) and the normality assumption together allow us to derive the log-likelihood function. The individual likelihood for any student i is given by

$$\begin{aligned} L_i(\Omega) =& f(\eta_{i,j,g,s,03-04}|\gamma^e, \theta, \rho, \mu, \sigma) \\ =& \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big\{ \frac{-1}{2\sigma^2} \cdot (y_{i,j,g,s,03-04} - y^C_{i,j,g,s,03-04} - \gamma_1^e e^N(\pi_{i,02-03}) \\ & - \theta e^N(\pi_{i,03-04}) - \rho \bar{e}^N_{s,02-03})^2 \Big\}. \end{aligned} \tag{D.6}$$

Taking the natural log and summing over all students in the state results in the following log-likelihood function:

$$\begin{aligned} l(\Omega) =& \sum_{i=1}^{N} log L_i(\Omega) \\ =& -\frac{N}{2} log(2\pi) - \frac{N}{2} log(\sigma^2) \\ & - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \Big(y_{i,j,g,s,03-04} - y^C_{i,j,g,s,03-04} - \gamma_1^e e^N(\pi_{i,02-03}) \\ & - \theta e^N(\pi_{i,03-04}) - \rho \bar{e}^N_{s,02-03} \Big)^2, \end{aligned} \tag{D.7}$$

where the full parameter vector is given by $\Omega = [\gamma_1^e, \theta, \rho, \mu, \sigma^2]'$.

Key to our estimation approach is the notion that while student-specific effort values in 2003-04 are

unknown to the econometrician, the function by which they are determined *is* known (under Assumption 5, justified above). When searching over values of γ_1^e to maximize the log likelihood, we can therefore use standard gradient-based methods by taking the derivative of the known effort function, $e^N(\cdot)$, with respect to γ_1^e. Specifically, on each iteration of the search, the routine selects a value for γ_1^e, substitutes that value (along with the other known inputs) into the known effort function $e^N(y^C_{i,j,g,03-04} + \gamma_1^e e_{j(i,02-03)} - y_g^{T,N})$ and generates an effort level in 2003-04 for each student. The iterative search continues until the routine arrives at a value for γ_1^e that, together with the 2003-04 effort levels it implies, maximizes the log likelihood.[56]

D.IV.ii Identification

In Section VI.B.3 of the main text, we claimed that any non-monotonic effort function in 2002-2003 and non-flat function in 2003-04 are sufficient for separate identification of γ_1^e and θ. To see why, suppose that the 2003-04 effort response is determined by some arbitrary non-flat function and 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 2002-2003, due to the non-monotonic profile of the effort function in that year. Incentive strength in 2003-04 shifts rightward for each student by the common amount of 2002-03 effort that persists. A non-flat effort function in 2003-04 then guarantees that at least some student pair satisfying the identical-effort condition in 2002-03 receives divergent levels of effort in 2003-04. Indeed, there is zero variation in effort within all such student pairs only if the 2003-04 effort function is flat, implying that any non-monotonic effort function in 2002-2003 and non-flat function in 2003-04 are sufficient for identification.

[56] In practice, we perform the estimation in MATLAB using the 'fmincon' command and supplying the gradient vector.

Appendix E Policy Analysis: Technical Details

E.I Estimating the Per-Teacher Monetary Equivalent of the NCLB Sanction

In this section, we provide supplemental details regarding our four-step strategy for using the implicit link between ABCs and NCLB incentives to estimate a monetary-equivalent value for the NCLB sanction.

Step 1 - NCLB's Effect on Subsequent ABCs Financial Incentives

We first 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. We then combine the difference in these passing probabilities with the ABCs per-teacher bonus payment of $750 to determine the expected dollar value each school stood to lose by responding to NCLB's introduction. To obtain the corresponding change in financial incentives for a one-unit change in NCLB school-level effort, we then regress the expected dollar value each school stood to lose in 2003-04 on average school-level effort from the prior year. We now explain this approach in more detail.

Calculating Schools' Expected Financial Losses Under the ABCs. To calculate the probability of passing the ABCs in 2003-04 assuming that NCLB was *not* enacted in 2002-03, recall from equation (D.4) that a school passes the ABCs in 2003-04 when the following condition is satisfied:

$$\sum_{g=3}^{G_s}\left(\left(\bar{y}^{C}_{g,s,03-04}-\alpha\bar{\hat{y}}_{g-1,s,02-03}\right)+\bar{e}_{g,03-04}+(\gamma_1^e-\alpha)\bar{e}_{g-1,02-03}+\bar{\eta}_{g,s,03-04}-\alpha\bar{\epsilon}_{g-1,s,02-03}\right)\geq 0. \tag{E.1}$$

Using this equation, we calculate school-level growth scores under the ABCs in 2003-04 for each school under the counterfactual scenario in which NCLB was not enacted, following the aggregation rules set out under the ABCs and substituting in values from our framework where appropriate. In particular, we set NCLB effort in 2002-03 and 2003-04 equal to zero ($\bar{e}_{g,03-04}=\bar{e}_{g-1,02-03}=0$). The predicted score in 2002-03 ($\hat{y}_{i,j',g-1,02-03}$) is thus the test score that would have arisen in that year without NCLB and the counterfactual predicted score from 2003-04 ($y^{C}_{ij,g,s,03-04}$) is the test score that would have occurred in 2003-04.

The differences between ABCs targets (which use lagged performance from 2002-03) and average student test scores are then used to form school-level ABCs growth scores.[57] In terms of our framework,

[57] While our framework 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.

a school passes the ABCs in 2003-04 under this scenario when

$$\sum_{g=3}^{G_s}\left(\left(\bar{y}^C_{g,s,03-04}-\alpha\bar{\hat{y}}_{g-1,02-03}\right)+\bar{\eta}_{g,03-04}\right)\geq 0. \tag{E.2}$$

We also calculate school-level growth scores under the ABCs in 2003-04 under a scenario in which schools only respond with additional effort in 2002-03. This is because we are interested in isolating the reduction in ABCs passing probabilities caused by the initial response to NCLB, so we do not incorporate 2003-04 NCLB effort responses into the calculation (by setting $\bar{e}_{g,03-04}=0$). 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 2003-04 to be the sum of the counterfactual predicted score and the persistence of the effort response from 2002-03. We again calculate the differences between 2003-04 performance and ABCs targets, using these differences to form school-level ABCs growth scores.[58] A school passes the ABCs in 2003-04 in this scenario when

$$\sum_{g=3}^{G_s}\left(\bar{y}^C_{g,s,03-04}+\gamma_1^e\bar{e}_{g-1,02-03}+\bar{\eta}_{g,03-04}-\alpha\bar{y}_{g-1,02-03}\right)\geq 0. \tag{E.3}$$

With school-specific ABCs growth scores, we are able to calculate school-level probabilities of passing the ABCs in 2003-04. To do so, we assume that average test score noise in each school ($\sum_{g=3}^{G_s}\bar{\eta}_{g,03-04}$) is distributed according to the cumulative density function $F(\cdot)$. We represent $F(\cdot)$ using a normal distribution with mean zero and assess the sensitivity of our analysis to a variety of alternatives for the standard deviation of this distribution. Specifically, we let the SD of school-level randomness vary from 0.1 to 1 developmental scale points in increments of 0.1.[59] In each case, we calculate the school-level probability of passing the ABCs assuming that NCLB was never enacted, the probability of passing the ABCs with only the 2002-03 NCLB response, and the difference between the two, reflecting the degree to which each school lowered the likelihood of passing because of its effort response to NCLB:

$$\Delta F_s = F\left(\sum_{g=3}^{G_s}\left(\bar{y}^C_{g,s,03-04}-\alpha\bar{\hat{y}}_{g-1,02-03}\right)\right)-F\left(\sum_{g=3}^{G_s}\left(\bar{y}^C_{g,s,03-04}+\gamma_1^e\bar{e}_{g-1,02-03}-\alpha\bar{y}_{g-1,02-03}\right)\right). \tag{E.4}$$

[58] 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 2003-04 predicted reading outcomes in the calculations. ABCs reading targets depend on both prior mathematics and reading scores, however, so despite using realized reading scores as 2003-04 outcomes in both scenarios, we do change the ABCs reading targets to incorporate prior counterfactual mathematics scores where appropriate.

[59] 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. Using even smaller values that are between 0.01 and 0.1 for the standard deviation does not alter any of our conclusions.

Panel (a) of Table E.1 provides a summary of the school-level passing probabilities under each counterfactual scenario. For a 0.1 scale-point standard deviation in noise, the average difference (across all schools) between the two passing probabilities is 20 percentage points, while it is 8 percentage points for a 1 scale-point standard deviation of noise, and monotonically decreasing in between. By responding to NCLB in 2002-03, the average school lowered its chances of passing the ABCs in 2003-04 by between 8 and 20 percentage points. Multiplying these figures by the ABCs bonus payment of \$750 implies that, the average school stood to lose between \$60 and \$150 per teacher in 2003-04 because of its effort response in 2002-03.

The NCLB effort response decreased the likelihood of ABCs target attainment the following year because the persistence rate of effort (γ_1^e) is lower than the required rate of growth under the ABCs, given by α.[60] For a given change in effort in 2002-03, the persistence of effort determines the rate at which the test score increases the following year, while the ABCs coefficient determines the rate at which the target increases. The discrepancy between the two parameters implies that the ABCs target increased at a faster rate than student test scores, thus making it more difficult for schools to pass the ABCs.[61]

Calculating the Change in Financial Incentives for a One-Unit Change in School Effort. To obtain the change in 2003-04 financial incentives in terms of a one-unit change in average school-level effort from the prior year, we regress the change in school-level ABCs passing probabilities in 2003-04 (given by equation (E.4)) on average school-level effort in the prior year:

$$\Delta F_s = \alpha + \beta \bar{e}^N_{s,02-03} + \nu_{s,03-04}. \tag{E.5}$$

The estimate $\hat{\beta} = \frac{d(\Delta F_s)}{d\bar{e}^N_{s,02-03}}$ governs the magnitude by which a one-unit increase in school-level effort in 2002-03 lowers the likelihood of ABCs target attainment in 2003-04. Panel (b) of Table E.1 reports the estimated coefficients from equation (E.5). The coefficient ranges from -0.32, when the standard deviation of school-level noise is assumed to be 0.1, to -0.09, when the standard deviation is assumed to be 1. A one-unit increase in average school-level effort in 2002-03 thus reduces the probability of passing the ABCs by a value between 9 and 32 percentage points, which amounts to between \$67.5 and \$240 per teacher when multiplied by the ABCs bonus payment of \$750.

Step 2 – NCLB's Effect on Subsequent Student Test Scores

In Section VI, we recovered the effect of a one-unit change in average school-level effort from 2002-03 on test scores in 2003-04 as $\hat{\rho} = \frac{dy}{d\bar{e}^N_{s,02-03}} = 0.29$. To aid with the exposition below, we now write $\hat{\rho}$ explicitly

[60] We estimate $\gamma_1^e = 0.10$ while the implied coefficient on the prior mathematics score under the ABCs is far higher, given by $\alpha = 0.68$.

[61] This is similar to the prediction in Macartney (2016).

as the effect of effort on scores, $\hat{\rho} = \frac{dy}{d\bar{e}^N_{s,02-03}}$. In the main text, we argued that this estimate reflects the *indirect* relationship between financial incentives under the ABCs and the corresponding teacher effort responses in 2003-04.

Step 3 – Connecting Changes in ABCs Financial Incentives and Test Scores

We now scale the effect of lagged school-level NCLB effort on test scores ($\frac{dy}{d\bar{e}^N_{s,02-03}} = \hat{\rho} = 0.29$) by the effect of lagged school-level NCLB effort on ABCs passing probabilities in 2003-2004 ($\hat{\beta} = \frac{d(\Delta F_s)}{d\bar{e}^N_{s,02-03}}$). Multiplying the result by the inverse of the ABCs bonus payment of \$750 results in the *direct* effect of financial incentives on test scores, $\lambda = \frac{\hat{\rho}}{750 \cdot \hat{\beta}} = \frac{dy}{750 \cdot d(\Delta F_s)}$, which we assume is driven by teacher effort. This calculation implies that a \$1 increase in financial incentives causes an effort-driven increase in student test scores that is between 0.0012 and 0.0043 developmental scale points. In terms of magnitude, these estimates imply that when teachers stand to lose \$375 in expectation,[62] they increase effort by between 0.05 and 0.16 (student-level test score) standard deviations.

Step 4 – Inferring the Value of the NCLB Sanction through the ABCs Link

The average school-level NCLB effort response in 2002-03 ($\bar{e}^N_{s,02-03}$) is 1.97 developmental scale points. Combining this average effort response under NCLB with our estimates of how effort responds to financial incentives implies that the NCLB sanction is valued between \$458 (1.97/0.0043) and \$1,640 (1.97/0.0012) per-teacher, where the range reflects the range of values for $\lambda = \frac{\hat{\rho}}{750 \cdot \hat{\beta}} = \frac{dy}{750 \cdot d(\Delta F_s)}$, which we obtain in Step 3 above.

E.II The Student Achievement Benefits of Incentive-Based Reforms

In this section, we explain how the incentive-based reform proposed in Section VII can generate a performance improvement among teachers equivalent to the performance improvement generated by the ability-based reform. The ability-based reform improves performance by two standard deviations of teacher ability for teachers in the bottom five percent of the ability distribution. We use this two standard-deviation improvement as our target, showing that an incentive-based reform can generate the same improvement by (i) using student-specific proficiency targets and (ii) adjusting the value of the monetary equivalent of the NCLB sanction.

As an overview, we first demonstrate that using student-specific proficiency targets to make all students marginal generates a teacher effort response of 0.29 (student-level) standard deviations of the test score, equivalent to 1.6 standard deviations of teacher ability. We then show that setting the value of the NCLB sanction to 25 percent of its current value causes the required additional 0.4 standard-deviation

[62] This is the expected amount when there is a 50 percent chance of satisfying the ABCs high-growth target.

effort response, making the total effect of the proposed reform equivalent to that of the ability-based reform.

Consider using student-specific proficiency targets to increase the fraction of marginal students in each classroom to 100 percent. Using the estimated relationships between marginal student presence and teacher performance in Panel (a) of Table 3, along with the corresponding numbers of students in each grade, we calculate the average effort response across all grades of such a policy as $\frac{2144}{7312}(1.55)(1) + \frac{2598}{7312}(4.39)(1) + \frac{2570}{7312}(2.41)(1) = 2.86$ 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 teacher effort response of 0.29 (or $\frac{2.86}{10}$) student-level test score standard deviations, equivalent to 1.6 standard deviations of teacher ability.

To generate the remaining 0.4 standard-deviation effort response, thus achieving the required full two standard-deviation performance change, we first set the value of the NCLB sanction equal to 125 percent (i.e., 2/1.6) of the current value. Central to this calculation is the assumption that teacher effort increases linearly with the value of the NCLB sanction.[63] It is worth noting, however, that the two-standard deviation effort response under the incentive-based reform is realized throughout the *entire* distribution of teachers while the benefits under the ability-based reform are realized only within the subset of the bottom five percent of teachers (in terms of value-added). Because of this fact, we place the incentive-based reform on a common footing by multiplying the 125 percent sanction value by 0.05 (i.e., 5 percent) to arrive at a sanction worth only 6 percent of its current value.

The 6 percent number 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. Yet our estimates of the persistence rates of teacher effort and ability indicate that effort decays faster than ability, implying that the benefits of the incentive-based reform fade out faster in the longer run. We therefore need to adjust the 6 percent figure to reflect the differential long-run persistence of effort and ability. The long-run comparison depends on how effort effects persist beyond one year.[64] 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 4.63 percent (i.e., 0.10*0.463) of the initial effort effect. Thus four periods forward, the incentive reform achieves 24 percent of the effect of the ability reform (i.e., 0.0463/0.19). Scaling the 6 percent figure by 24 percent (i.e., 0.06/0.24) thus results in a sanction valued at 25 percent of its current value.

[63] We discuss the implications of relaxing this assumption in Section VII.D of the main text.

[64] This is an issue we intend to investigate in future work.

Table E.1: 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	0.72	0.70	0.67	0.65	0.63	0.62	0.60	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	-0.20	-0.19	-0.18	-0.15	-0.14	-0.12	-0.11	-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	-0.32 (0.04)	-0.27 (0.03)	-0.22 (0.02)	-0.18 (0.02)	-0.16 (0.02)	-0.13 (0.01)	-0.12 (0.01)	-0.10 (0.01)	-0.09 (0.01)	-0.09 (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 2003-04. In panel (a), we calculate the probability of passing the ABCs for each school, assuming NCLB only operated in 2002-03 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 2002-03, 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.