Teacher Value-Added and Economic Agency*

by
Hugh Macartney, Robert McMillan, and Uros Petronijevic†

December 2019

Abstract

We present an estimable framework expressing teacher value-added in terms of teacher ability and effort, along with a strategy to identify these two unobserved inputs and their distinct test score effects for the first time. This uses exogenous incentive policy variation and rich longitudinal data from North Carolina. We find both inputs raise current and future scores, and effort responds systematically to incentives, underlining the agency of the teaching force. To explore the policy implications of this agency, we use our estimates to compare the cost effectiveness of incentive and ability-based education reforms, finding incentive reforms often come out ahead.

Keywords: Incentives, Education Production, Effort, Ability, Teacher Value-Added, Accountability, Education Policy, Cost-Effectiveness, Persistence

*This is an updated version of NBER Working Paper 24747. We would like to thank Raj Chetty, Damon Clark, Peter Cziraki, John Friedman, Elaine Guo, Caroline Hoxby, Magne Mogstad, Louis-Philippe Morin, Jonah Rockoff, Juan Carlos Suárez Serrato, Hammad Shaikh, Brooklynn Zhu, and seminar participants at Arizona State University, Chicago Harris, Columbia University, Duke University, McMaster University, Stockholm 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.

†Contact information: Macartney – Duke University and NBER, hugh.macartney@duke.edu; McMillan – University of Toronto and NBER, mcmillan@chass.utoronto.ca; Petronijevic – York University, upetroni@yorku.ca.

I Introduction

In the quest for viable policies to improve public school outcomes, considerable attention has focused on teacher performance, informed by the ever-wider availability of detailed administrative data sets that link students to their teachers. Leveraging such data, sophisticated teacher value-added measures have been developed in an influential recent literature, intended to capture the overall performance impact of a given teacher – see, for example, Kane and Staiger (2008) and Chetty, Friedman and Rockoff (2014a,b).[1] In turn, estimates of these value-added measures are now the cornerstone of policy interventions that include, somewhat controversially, dismissing low value-added teachers.

Teacher value-added ('VA') is not a primitive of the education production technology.[2] Instead, it is natural to attribute teacher performance to a teacher's 'type' (or ability) and actions (or work effort), with effort likely to depend on the incentive environment.[3] Yet teacher ability and effort are typically not observed, and partly as a consequence, we do not have estimates of the two education inputs or their longer-term impacts on a common footing. This in turn inhibits policy design, making it difficult to compare the cost effectiveness of reforms that target teacher effort versus teacher ability in any systematic way.

Motivated by these estimation and policy issues, our paper has two goals. On the estimation front, we seek to develop an approach that allows both incentive-varying teacher effort and incentive-invariant teacher ability to be identified along with their persistent effects for the first time in the literature.[4] The resulting estimates can provide insight into the respective short- and longer-term productivities of these education inputs, filling a gap in our knowledge of the education process. They also have a bearing on the policy debate, as different reforms are likely to be suitable for bringing forth different inputs. On that policy theme, our second goal is to set out a quantitative means for comparing incentive reforms with relevant alternative education policies, including policies that target teacher ability, drawing on the new estimates; we will do

[1] The value-added literature builds on well-known studies – notably Rockoff (2004) and Rivkin, Hanushek and Kain (2005) – using fixed effects methods to show convincingly that 'teacher quality' matters.

[2] Consistent with this claim, recent empirical research shows clearly how teacher performance varies with workplace characteristics – see Jackson and Bruegemann (2009) and Jackson (2013), for example. (Appendix A provides a fuller review of this and other related literatures.)

[3] This accords with evidence that accountability reforms in education have improved student achievement in many settings – see Figlio and Loeb (2011) for a comprehensive survey.

[4] We take 'effort' to be any incentive-related action that raises scores – for instance, exerting more effort in the classroom or devoting more time to lesson planning. We will define ability as the component of teacher VA that does not change over time, conditional on teacher experience.

so in terms of their performance benefits and financial costs. This formal policy comparison is novel, and should help to sharpen policy makers' assessment of viable education reform options.

Central to addressing the first goal is an estimable framework built around an education production function, which we use to express teacher VA in terms of teacher ability and effort. As the form of the production technology is unknown, we assume a linear approximation to the true function and let inputs have persistent impacts over time – appropriate given that knowledge accumulates in an education context. From this specification of the technology, we derive estimating equations that guide our empirical approach.

To bring the framework to the data, we take advantage of a rich administrative data set covering all North Carolina public school teachers and students over time, and exogenous incentive 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) and Neal and Schanzenbach (2010), for instance – proficiency schemes such as NCLB make students matter differentially depending on how marginal they are. In line with this idea, we show as motivation for the main analysis that teacher VA *increases* in an intuitive measure of incentive strength, a finding robust to using across- and within-teacher variation, with the evidence also pointing to a teacher effort response to the reform's introduction.[5]

Our proposed estimation procedure leverages this accountability policy variation. Specifically, we assume that educators respond to incentives under proficiency systems by directing relatively more effort to students predicted to score near the proficiency threshold (as the evidence in our application indicates) and in a similar way across years. We estimate the contemporaneous ability of teachers observed before and after the reform using standard VA estimation methods and pre-reform data. We then identify contemporaneous effort from the correlation of post-reform VA with incentive strength (as 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.

Next, we investigate the extent to which these contemporaneous teacher ability and effort measures persist in determining an individual 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

[5]The reform did not cause school principals to engage in within-school teacher reassignments nor to alter class sizes – leading alternative explanations.

method from the previous literature, finding that approximately 40 percent of the initial impact of teacher ability persists after one year, and 20 percent after four years.

To identify the persistence of effort, the ideal experiment would involve a single one-time incentive reform that resulted in an immediate effort response, with no correlated responses in future, allowing us to use a strategy similar to the approach for estimating ability persistence. As there are several departures from that ideal in our observational setting,[6] we develop a maximum likelihood estimation procedure based on the technology that allows us to account for three distinct sources of teacher effort explicitly, including the persistent effects of effort from the previous year (see Section V for a detailed description). The resulting estimates indicate that 10 percent of the initial effort effect persists one year ahead, which amounts to 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 rarely identified empirically. Further, we show that not accounting for the test score effects of contemporaneous effort decisions would result in a significant overestimate of effort persistence.

The combined estimation approach for recovering teacher ability, effort, and their persistent effects – the first main contribution of our study – is applicable in other contexts where rich administrative data and clear policy variation are available. Of note, the new estimates we obtain show that teacher effort is both a productive input and one that responds systematically to incentive variation, with longer-term benefits for students: in essence, teacher actions matter alongside teacher type. Our analysis thus underscores the *agency* of the teaching force as an important feature of the education production process and one that can be influenced by the incentive environment, itself partly under the control of the policy maker.

As our second main contribution, we explore the quantitative implications of this economic agency for the education policy debate. Our estimates suggest a natural question: For resource-constrained decision makers, are policies that alter the mix of teachers based on value-added preferable to policies that take the existing teacher stock as given and alter effort incentives instead? In particular, how would the widely discussed policy of replacing the bottom five percent of the teacher value-added distribution compare with a policy that targeted teacher effort incentives? Our framework and estimates allow us to provide the first quantitative evidence

[6]These include the correlation of NCLB effort incentives over time, the dependence of contemporaneous effort on the persistence parameter we wish to estimate, and induced changes to a concurrent state-level accountability program – the ABCs of Public Education – following the introduction of NCLB.

addressing this relevant question.

For the costs and benefits of the ability-based policy, we appeal to prior research – see Hanushek (2009, 2011), Chetty *et al.* (2014b) and Rothstein (2015) – for plausible estimates of the likely impacts in our setting. The literature suggests that a policy removing the bottom five percent of the teacher VA distribution would cost around $700 per teacher, not least by increasing employment risk. At the same time, the policy would be likely to increase long run test scores by 0.0034 SD (averaging across all teachers), drawing on our estimates of the persistence of ability.

The viable incentive policy we consider is based on the state's pre-existing ABCs accountability system. This provides incentives to improve student performance throughout the distribution, setting individualized targets for all students according to their prior performance and thus making all students marginal. To calculate the benefits and costs of a directly comparable incentive policy, we need to recover the underlying mapping between output and incentive expenditures. Here, we use the introduction of NCLB and the exogenous variation it produced in both (i) the incentive cost of increased effort[7] and (ii) the response of output to effort (due to the heightened incentives induced under the new ABCs targets the following year). We show how to combine the two to recover the desired output-expenditure mapping (see Section VI). This in hand, we can adjust the rewards under the incentive scheme to equate costs with the ability reform, and then compare the resulting output.

We find that the incentive reform produces 88 percent higher output for the same cost as the ability-based reform, using our preferred estimates. Further, we show that the advantage enjoyed by the incentive reform is likely to arise in a range of plausible circumstances – a robust advantage stemming from the fact that incentives can be applied throughout the teacher VA distribution whereas ability-based reforms focus on part of the distribution. Not only does the policy analysis indicate that incentive reforms are competitive with policies targeting teacher ability. It shows, more generally, how incentive reforms can be compared with alternatives in a systematic way that draws on underlying technology estimates, providing policy makers with a new means to choose among the menu of viable education policy options.

The remainder of the paper is organized as follows: Section II sets out the education produc-

[7]The effort response to NCLB, which we have documented, raised the targets that schools faced in the following year, given the way ABCs targets are set. More demanding targets in turn lowered the probability of achieving the ABCs bonus. We can monetize the cost of extra effort from this change in probability.

tion technology central to the analysis. Section III describes the accountability programs and the North Carolina administrative data we use for estimation, along with motivating descriptive evidence. In Section IV, we present our strategy for recovering ability and effort contemporaneously, followed by the results from that exercise, and in Section V, we describe our approach for estimating the persistent effects of ability and effort along with the associated findings. Section VI presents the cost-effectiveness analysis, and Section VII concludes.

II Framework

This section sets out the conceptual framework at the heart of our study. We use it to state the specific goals of our econometric analysis and to derive the equations that guide the estimation approach we develop.

The framework is based on an education production technology that relates inputs to measured education output y. 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 is to impose some minimal structure (which we do in this section), then leverage plausible sources of policy variation and rich longitudinal data, described in Section III, to identify key inputs.

We make two main assumptions about the technology's form:

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

This serves as an approximation to the true underlying function.

Assumption 2: Inputs have a cumulative effect on output.

This second assumption is natural 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. Specifically, 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 focus on two inputs, teacher ability and teacher effort, and their contemporaneous and persistent effects in terms of measured output given by test scores. Reflecting the two

assumptions above, the following representation of the technology makes explicit (suppressing other inputs for clarity) how each of the two inputs affects student learning, both in the current year t and in all prior academic years:

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

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.[8] 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 $t-g$), where τ is a 'lag' index that takes on integer values from 0 up to 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, and is allowed to be student-specific (as with NCLB), and ν_{ijgst} 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. We also impose the following

Normalization: The contemporaneous effects of a one-unit change in teacher ability and a one-unit change in teacher effort are normalized to be equal, with $\gamma_0^a = \gamma_0^e = 1$.

We make this normalization because the unobserved contemporaneous inputs and parameters cannot be separately identified, although potential differences in the persistent effects of ability and effort are still allowed – that is, $\gamma_\tau^a \neq \gamma_\tau^e$ for $\tau > 0$.

In an ideal setting where equation (1) could be estimated directly, it would be possible to identify teacher ability and effort, $a_{j(i,t-\tau)}$ and $e_{j(i,t-\tau)}$, separately for each teacher j, 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 almost never possible to observe the full sequence of the relevant ability and effort inputs that students have received since the start of their formal schooling. As a consequence, our approach will be to summarize inputs from the more distant past using the lagged test score of a given student, an approach used widely in the literature – see Rivkin *et al.* (2005), for example.

[8] We take teacher and student assignments to classrooms as given here, noting that the empirics below will address non-random sorting.

Second, and related, identifying the full sequence of persistence parameters is not feasible, so we will concentrate on identifying a subset.

Our proposed estimation strategies will be built around a re-writing of the test score technology in terms of once-lagged test scores. After some straightforward algebra,[9] the resulting expression for current test scores is then

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

where the error term ϵ_{ijgst} contains the entire history of past inputs as well as the two most recent random shocks to performance.[10]

The first empirical goal of the paper is to separate out the contemporaneous teacher ability and effort inputs. Equation (3) will guide our approach (presented in Section IV) for estimating contemporaneous ability and effort at the *teacher* level, using estimates of teacher VA as a starting point. Appealing to the additive structure, teacher VA in a given year is captured by a teacher-year fixed effect q_{jt}, written $q_{jt} = a_j + \bar{e}_{jt} + \bar{\epsilon}_{jt}$, consisting 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 ($\bar{\epsilon}_{jt}$).

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

[9] First, multiply the prior score by γ, which represents the rate at which the stock of knowledge accumulated up to period $t-1$ persists to affect current test scores (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, allowing the teacher j' and school s' in the previous year to be different and imposing the above normalization. Doing so yields

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

Then move the lagged score back to the RHS and relabel.

[10] From (2), the error ϵ_{ijgst} is given by $\sum_{0\le\tau\le g-1}\left[(\gamma^a_\tau - \gamma\gamma^a_{\tau-1})a_{j(i,t-\tau)} + (\gamma^e_\tau - \gamma\gamma^e_{\tau-1})e_{j(i,t-\tau)}\right] + (\nu_{ijgst} - \gamma\nu_{i,j',g-1,s',t-1})$. When teacher ability and effort both decay at a common rate, γ, the error term consists only of random performance shocks, given by $\nu_{ijgst} - \gamma\nu_{i,j',g-1,s',t-1}$.

can be credibly identified, we re-write equation (3),[11] and express the production technology as

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

The second goal with reference to equation (4) is then to identify the persistent effects of teacher ability and effort one year ahead, given by γ_1^a and γ_1^e. This specification will serve as the basis for the estimation approach we implement in Section V.

III Institutional Background and Data

Our broad estimation approach requires exogenous incentive variation and rich data on individual students and teachers. Given those needs, we focus on North Carolina, a state that offers useful variation in performance incentives across teachers and schools, as well as administrative data covering all public schools and their teachers and students, followed over time.

III.A Accountability Incentives

Incentive variation in the state arises from two separate accountability regimes. NCLB 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 emphasized student proficiency, establishing performance targets based on end-of-grade mathematics and reading tests. When a smaller-than-required fraction of students reached proficiency status on those tests, schools failed to satisfy NCLB requirements and were subject to sanctions that became more severe over time in the event of repeated failure. As is well appreciated – see Reback (2008), for example – such proficiency-count systems create incentives for teachers to direct relatively more effort toward students likely to score close to the test score proficiency target.

NCLB was introduced on top of a pre-existing state-level accountability program, the ABCs of Public Education, implemented in North Carolina 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

[11] Specifically, 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}, denote the new error term η_{ijgst}, and use the normalization $\gamma_0^a = \gamma_0^e = 1$.

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 – a feature we use in the cost-effectiveness analysis below.[12]

It is worth emphasizing the marked shift in incentives facing schools and teachers in North Carolina that followed the introduction of NCLB in the 2002-03 academic year. The ABCs' emphasis on average performance meant that teachers were not previously incentivized to direct effort across individual students in a differential way. Yet because of NCLB's sole focus on proficiency, it is reasonable to expect that students likely to score near the test score proficiency threshold in 2002-03 would realize greater test score gains than in past years, given the new incentives. We provide evidence of this pattern below, and use it to motivate our strategy for separately identifying teacher effort and ability.[13]

At the same time, in terms of possible confounding changes, the introduction of NCLB was *not* associated with the creation of new end-of-grade tests or student proficiency thresholds based on those tests – we discuss this in more detail in Section IV. Rather, the state already had evaluations and performance metrics in place as part of the ABCs.

III.B Data and Descriptive Statistics

The rich longitudinal data used in our analysis cover 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 over time, and linked to a teacher and school in any given year.

We provide an overview of the data here, with more information in Appendix B. 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 presents summary statistics. In terms of performance measures, we focus on end-of-grade test scores for students in third through fifth grade: the teacher recorded as the test proctor in these grades is typically the teacher who taught the students throughout the year. The 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

[12] See Macartney (2016) for additional detail about the ABCs program.

[13] Given the complexity of the NCLB legislation, other plausible sources of variation are available to researchers – for instance, relating to subgroups. We focus on student-level variation associated with a straightforward measure of incentive strength (introduced in Section III.C), showing this to be first-order.

grade (see Appendix B for further discussion). 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.

We work directly with test scores measured on the state-provided developmental scale rather than standardizing test scores at the grade-year level (as in much of the literature) because of our focus on the incentive effects associated with NCLB. This dictates that we preserve as much of incentive-related variation in the data as possible – in particular, the aggregate effects of NCLB incentives in the 2002-03 academic year,[14] and also consider the same test score metric teachers are likely using to make effort decisions.[15] As with many education studies that use test scores, our analysis will draw on the careful psychometric design of such tests (in North Carolina and elsewhere).[16] In this regard, compelling prior papers have already demonstrated the strong 'signal' value contained in test scores, apparent in terms of their correlation with important long-run outcomes (see, for example, CFR 2014b).

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. As the table shows, mathematics and reading growth is positive on average across grades, with the largest gains in both subjects occurring in the earlier grades.[17]

The data set also includes demographic characteristics shown in the table 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.

[14] De-meaning test scores is typically carried out precisely in order to *remove* the effects of year-specific influences on performance.

[15] North Carolina embedded test score proficiency cutoffs into the developmental scale directly, which makes it natural to assume that schools and teachers used the same scale when forecasting which students were likely to score near the test score proficiency threshold. This consideration will become especially important when we present our strategy for estimating the persistence of teacher effort in Section V.

[16] We note that studies that standardize test scores rely – at least implicitly – on the integrity of the underlying developmental scales used to measure student performance, given they are order-preserving linear transformations of developmental scale test scores that are not invariant to the original scale.

[17] 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 eighth grades, which are used when measuring the persistent effects of teacher ability and effort below.

III.C Descriptive Evidence

We now motivate the subsequent empirical analysis with correlations that suggest teacher value-added is responsive to incentives.

Proficiency-count systems like NCLB (as noted) provide educators with strong incentives to focus on students at the margin of passing relative to the scheme's fixed proficiency target. Building on that notion, we show that teacher-year fixed effects,[18] which are commonly used to measure teacher effectiveness, covary with a simple proxy for NCLB incentive strength in 2002-03. Teacher-year VA should be an increasing function of classroom *average* student NCLB incentive strength, since teacher-year VA represents *average* residual student test score gains within a classroom. Thus 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 – the results are robust to various alternative choices (see Section IV).

Figure 1 shows that teacher-year fixed effects are positively correlated with the proportion of marginal students within a classroom. The relationships in Figure 1 are 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, we would expect there to be no relationship between the proportion of marginal students in a classroom and teacher VA in the pre-NCLB period: we will document such a pattern in the next section.[19]

IV Separating Contemporaneous Ability and Effort

Building on the descriptive evidence, we now set out our estimation approach for separating a teacher's contribution to student test scores (measured by VA) into *current* ability and effort, before presenting the estimates.

[18] Appendix C describes how teacher-year fixed effects are estimated, following standard methods.

[19] 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 we present in the next section will account for sorting.

IV.A Estimation Approach

Our estimation approach consists of three steps.

Step 1 - Estimating Teacher-Year Fixed Effects: We use standard methods to compute teacher-year fixed effects based on student mathematics scores for each 'teacher j and academic year t' combination.[20] We start by aggregating the individual production technology given by equation (3) up to the teacher level. In line with our framework, this allows us to write the estimated 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 will identify the ability and effort components in equation (5) separately in the next two steps.

Teacher 'ability' in this analysis should be thought of as capturing both (true) ability and average ABCs-related effort exerted by the teacher across all of her years of teaching under the ABCs program, given that it operated in North Carolina prior to NCLB (see Appendix C.III for further discussion): the two cannot be separately identified. Understanding 'ability' in this 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.

Step 2 - Estimating Incentive-Invariant Ability: Next, we identify teacher ability during a period when NCLB did not operate, based on pre-reform data; this ensures that our estimates of teacher ability are independent of performance variation due to NCLB incentives. We use 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.[21] Specifically, we run the following pooled regression across all grades and years

[20] The approach as well as the students and teachers in the VA estimation sample are described in Appendix C.

[21] We opt not to use an estimator that allows teacher ability to drift over time (see CFR (2014a), Rothstein (2014) and Bacher-Hicks *et al.* (2014)), given that predicting teacher ability in 2002-03 with the drift estimator requires performance data from that year when constructing optimal weights, and could confound teacher ability estimates with incentive variation in that year. Also, the main advantage of the drift estimator is that it assigns greater weight to more recent years in order to better predict teacher performance in a given year. While CFR

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:[22]

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

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}.[23]

Step 3 - Estimating NCLB-Induced Effort Response: In the third step, we estimate NCLB-induced teacher effort using the estimated teacher fixed effects from 2002-03 from Step 1, along with estimates of teacher ability from Step 2 and the fraction of students in a teacher's classroom deemed 'marginal' with respect to the NCLB target.[24]

We define a student as 'marginal' if she was predicted to score within four developmental scale points of the test score proficiency cutoff on either side (as in the descriptive analysis in Section III.C). For each classroom, the incentive strength measure, m_{jt}, is defined as the fraction of students in the classroom who are marginal in this sense. 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 controlling for incentive-invariant teacher ability ($\hat{a}_j$) and teacher experience (exp_{jt}):

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

The error term ξ_{jt} captures potentially confounding classroom shocks.

(2014a) show this improves teacher VA prediction in their large urban school district, the correlation in North Carolina between teacher effect measures over time is higher than in the CFR data (see Rothstein 2014), implying a smaller benefit to using the drift estimator in our setting.

[22]The experience function $h(\cdot)$ is parameterized by including indicators for each level of experience from zero to five years, the omitted category being teachers with six or more years of experience; we choose this specification to be consistent with CFR (2014a).

[23]While the EB estimator is our preferred method for measuring teacher ability, our estimates of the effects of NCLB incentives on teacher-year VA are robust to accounting for teacher incentive-invariant ability in a non-parametric way using either first-difference or teacher fixed-effects models (see Appendix D).

[24]This draws 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.

Once NCLB is introduced, teachers are hypothesized to exert additional effort according to the amount of incentive pressure they face. We thus test whether there is a systematic relationship between $\hat{q}_{jt}$ and m_{jt} in 2002-03 but no relationship prior. Given the linearity assumption, we use equation (7) to predict the portion of teacher performance given by effort, writing $e(m_{j,02-03}) \equiv \hat{\psi} m_{j,02-03}$.[25] 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.

The key assumption underpinning our strategy for identifying teacher effort responses to NCLB incentives is that, conditional on teacher ability and experience, no other factors influencing teacher-year VA are correlated with m_{jt} in 2002-03 – an assumption whose validity we assess below – see Section IV.B.1.

IV.B Estimates of Ability and Effort

We now present the estimates from applying the approach.

Ability: In terms of our estimates of contemporaneous teacher ability, Table 2 reports 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. Figure 2 presents the incentive-invariant teacher ability distributions, where incentive-invariant ability is defined as the EB estimate from equation (6). The figure shows that there is significant variation in ability across teachers, all centered roughly around zero (due to a normalization in the EB procedure).

Effort: For effort, we start by presenting the estimated relationships between teacher value-added and incentives. The panels of Figure 3 show the grade-specific partial relationships between the teacher-year fixed effects $\hat{q}_{jt}$ and incentive strength m_{jt}, where the latter is residualized with respect to teacher ability and experience. For each grade, we present plots for 2002-03, indicating a clear increasing relationship between the part of the teacher-year effect

[25] By omitting an intercept, this parameterization assumes that NCLB-related teacher effort is zero when a teacher has no marginal students in her classroom. This is supported by the motivating visual evidence – see the binned scatter plots of Figure 1 and in Figures 3 and D.2 below. Our main results below focus on the variance of predicted effort and marginal changes in effort, both of which are determined solely by the slope parameter, $\hat{\psi}$.

unexplained by ability and experience and the proportion of marginal students in the classroom. Alongside, we plot the corresponding pooled relationship for all pre-NCLB years that also includes year fixed effects. The pre-NCLB variation plots show no discernible link between teacher performance and our measure of NCLB incentives, results that are robust to alternative cutoff points for defining a student as marginal under NCLB (see Appendix D).

Table 3 shows regression estimates indicating how teacher-year effects change with an increase in the fraction of marginal students in the classroom (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 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.[26]

The panels of Figure 4 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 for 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.

IV.B.1 Robustness Checks: Assessing Rival Hypotheses

The evidence just presented 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 potential explanations. We consider several hypotheses that are alternatives to teacher effort setting, summarizing our findings here – see Appendix D for the results in full.

First, our results are not explained by other institutional changes coinciding with NCLB's introduction. The state did not change either its curriculum or the content appearing on the end-

[26]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).

of-grade tests we use as our dependent variable,[27] nor did NCLB's introduction in North Carolina provide any new information to families about student performance, thus making parental or student effort responses unlikely. Second, we rule out several other school-level adjustments to NCLB as drivers of our results, showing that marginal students were not sorted differentially to classrooms based on teacher ability. Further, accounting for changes in class size and several other classroom-level student characteristics does not affect our main estimates.

V Estimating the Persistence of Teacher Ability and Effort

Having identified teacher ability and effort separately, we now assess whether these two inputs persist at different rates. The issue is key to understanding whether the separate effort effect we have identified is likely to be consequential for economic outcomes in the longer run, and also important for the policy analysis that follows. Unlike the previous section, the analysis is conducted at the student (rather than the teacher) level, thereby exploiting more variation in the data and following prior work that estimates the persistence of teacher effects as a whole.

V.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 the time index 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:

$$y_{i,j,g,s,t+n} = f_g(y_{i,j',g-1,s',t}) + x'_{ijgst}\beta + h(exp_{jt}) + \phi_n \hat{a_j} + \epsilon_{ijgst}. \tag{8}$$

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.[28] The coefficient ϕ_n represents the degree to which the effect of teacher ability from year t influences test scores in year $t+n$. Looking one period ahead, ϕ_1 is our empirical estimate of the persistence rate of teacher ability that is given by γ_1^a in equation (4), while setting $n=4$ lets us investigate the persistent effect

[27] The mathematics test was last changed in 2000-01.

[28] 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).

of teacher ability on test scores four periods ahead.[29]

Figure 5 presents the estimated ϕ_n coefficients from 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.[30] 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 $t+4$. (For reference, in their analysis of the persistence of teacher value-added, CFR (2014b) find that approximately half of the initial effect persists one period ahead and 20 percent remains four periods ahead.)

V.B Estimating the Persistence of Effort

Identifying effort persistence requires that we account for three confounding effort effects. First, incentives to exert effort under NCLB are strongly correlated over time – that is, students who are marginal in one year tend to be marginal the next and so would be expected to receive higher effort in both years as a result; thus it is important to control 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 will be determined in part by any effort persistence from the previous year; a strategy for disentangling the two is therefore needed. Third, the introduction of NCLB induced changes in incentives under the pre-existing ABCs program.[31] The resulting effort responses need to be controlled for to avoid confounding student-level effort persistence with school-level ABCs-related improvements.

The approach we devise involves estimating effort at the student level, focusing on effort

[29] Our structural parameter of interest, γ_1^a, reflects the direct effect of the prior-year teacher on current-year test scores. In contrast, the reduced-form estimate ϕ_1 represents the *total* effect of the prior-year teacher, which includes the direct effect and any potential effects that come from being tracked to better teachers in future years as a result of having a better teacher in the prior year. CFR (2014b) propose a strategy for estimating the direct (or net) effect of the prior-year teacher on students' adult earnings and show there are only modest differences between the net and reduced-form effects.

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

[31] Specifically, given the value-added nature of the ABCs system, higher performance in one year engenders higher targets the next, which will lead to stronger incentives to exert higher effort as a result.

persistence one year ahead, and using the production technology in a maximum likelihood procedure to account for the three effects just described. We now set out these elements (see Appendix E for a detailed description) before presenting the estimates.

V.B.1 Estimating Student-Level Effort

Under the proficiency-count design of NCLB, effort incentives can vary across students *within* a given classroom, with some students being more marginal than others: such within-classroom variation is entirely first order in our setting.[32] Thus we start by constructing a student-level measure of effort, rather than one that is common to all students taught by a given teacher.

Our effort measure is derived from the non-parametric patterns in Figure 6. This figure shows that the introduction of NCLB had pronounced non-linear effects on test scores, consistent with strong teacher effort responses to the scheme. Two notions are relevant for understanding the figure. First, we construct the *predicted student score* in 2002-03 in a way that excludes the NCLB-induced effort response,[33] as if NCLB had never been implemented in that year. The resulting difference between the realized and predicted scores for a given student then provides a (noisy) measure of the 2002-03 effort response by her teacher (where effort is taken to be the incentive-related boost to scores), used to construct the points on the vertical axis described below. Second, the *incentive strength* for each student is given by the distance between the predicted score and the fixed NCLB proficiency target. This incentive measure is used to form the horizontal axis in Figure 6, which groups students into two-scale-point width bins of incentive strength in 2002-03 (denoted $\pi_{i,02-03}$ for student i). We then plot the *average* difference between the realized and predicted score within each incentive strength bin on the vertical axis, 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 – 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 no NCLB effort response can occur) to ensure that we do not systematically under- or over-predict test scores at different parts of the distribution. Doing so makes clear that our predicted score tracks the

[32] In 2002-03, for example, fully 75 percent of the variance in the incentive strength measure we devise (see below) occurred within-classroom.

[33] Formally, $\hat{y}_{i,j,g,s,02-03} \equiv \gamma y_{i,j',g-1,s',01-02} + a_{j(i,02-03)}$ – see Definition D.1 in Appendix E.II.

realized score well throughout, given by the approximately flat line. In turn, this lends credence to the view that the 2002-03 pattern reflects student-specific NCLB effort.

The profiles for the two years in Figure 6 are then used to estimate an effort function that takes incentive strength as its argument. 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) – see Appendix E.III. The resulting effort function, denoted by $e^N(\cdot)$, is plotted in Figure 7.

We use this function to assign effort levels to individual students directly. Taking the student-specific values of $\pi_{i,02-03}$ and the function $e^N(\cdot)$, the effort dedicated to each student i by teacher j in 2002-03 is given by $e_{j(i,02-03)} = e^N(\pi_{i,02-03})$, reading off the appropriate effort level from the function.[34] (As explained below, we also use the function to assign a level of effort to each student in 2003-04, subject to further scaling.)

V.B.2 The Estimating Equation

Next, we specify an equation that can be taken to the data in order to estimate the rate at which effort persists along with other policy-relevant parameters. Here, we draw on the technology in equation (4) to obtain an expression for test scores in 2003-04 as a function of inputs in that year and inputs from the previous year whose effects persist:

$$\begin{aligned} y_{i,j,g,s,03-04} = \gamma(&y_{i,j',g-1,s,02-03} - a_{j(i,02-03)} - e_{j(i,02-03)}) \\ &+ \gamma_1^a a_{j(i,02-03)} + a_{j(i,03-04)} \\ &+ \gamma_1^e e_{j(i,02-03)} + e_{j(i,03-04)} + \eta_{i,j,g,s,03-04}. \end{aligned} \tag{9}$$

The RHS of this equation captures, on the first line, the persistent effect of once-lagged scores from 2002-03 excluding teacher ability or effort, written $y_{i,j',g-1,02-03} - a_{j(i,02-03)} - e_{j(i,02-03)}$, where the j-subscripts coincide – that is, $j' = j(i, 02-03)$. Given the technology, we interpret this component as the persistent effect of non-ability and non-effort inputs. The second line captures the persistent effects of teacher ability from 2002-03 and teacher ability in the current

[34] Note that the values of the effort function are negative for the left and right extremes of incentive strength. This should be interpreted as follows: the effort function reflects student test score gains *relative* to the pre-NCLB status quo; thus, 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 a subset of non-marginal students.

year 2003-04; and the third line includes the persistent effects of teacher effort from 2002-03 and in the current year 2003-04, along with a random shock to current test scores.

To derive the main estimating equation, we wish to isolate the effort components in 2003-04 that are relevant from an estimation perspective. Collect terms by defining $y^C_{i,j,g,s,03-04} \equiv \gamma(y_{i,j',g-1,s',02-03} - a_{j(i,02-03)} - e_{j(i,02-03)}) + a_{j(i,03-04)} + \gamma^a_1 a_{j(i,02-03)}$, and deduct this from both sides of (4), removing all non-effort inputs from the RHS. This yields our estimating equation, written:

$$y_{i,j,g,s,03-04} - y^C_{i,j,g,s,03-04} = \gamma^e_1 e^N(\pi_{i,02-03}) + \underbrace{\theta e^N(\pi_{i,03-04}) + \rho \bar{e}^N_{s,02-03}}_{e_{i,j,g,s,03-04}} + \nu_{i,j,g,s,03-04}. \qquad (10)$$

We discuss each of the three effort components on the RHS in turn, explaining where the corresponding effort incentives come from and the specific timing of each effect.[35]

The first term captures the effect of effort in the previous year, 2002-03, due to the introduction of NCLB incentives. We assume this effort component is known, determined by incentive strength ($\pi_{i,02-03}$) according to the semi-parametric effort function $e^N(\cdot)$. The parameter γ^e_1 measures the rate at which effort in 2002-03 persists one year ahead to affect test scores in 2003-04.

Next, the underbrace on the RHS of (10) serves to emphasize that in our formulation, the second and third terms are subcomponents of contemporaneous effort, $e_{i,j,g,s,03-04}$. Given the prevailing incentives in 2003-04, we hypothesize that contemporaneous effort comes from two sources. Taking these in turn, the second term (given by $\theta e^N(\pi_{i,03-04})$) consists of NCLB-induced effort in 2003-04. While it is unknown to the econometrician, we make the following 'shape' assumption:

Assumption 3: 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 $e^N(\cdot)$ as in 2002-03 to set effort, the 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. This formulation is reasonable given that the rules of NCLB operated in 2002-03 and 2003-04: it

[35]The reasoning is set out in full in Appendix E.III; for reference, a complete listing of all the formal notation used is given in Appendix Table E.1.

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, though possibly to a lesser or greater extent across years (scaled by θ).

The third term accounts for the test score effects of induced changes in effort incentives under the ABCs in 2003-04 (captured by the parameter ρ). ABCs incentives vary at the school level, not across students within a given school.[36] Because school ABCs targets are functions of student *average* prior-year test scores, they also depend on *average* prior-year effort (see Appendix E.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 subsequent ABCs targets. The fourth term on the RHS consists of random factors affecting individual scores in 2003-04.

V.C Estimation Procedure

Our goal is to recover the parameters of equation (10). These govern effort persistence (γ_1^e), the scale factor multiplying contemporaneous effort (θ), and the indirect effect of ABCs incentives on student test scores (ρ). One 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.

Our strategy involves estimating 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 (10):

Assumption 4 – Normality: $\eta_{i,j,g,s,03-04} \sim N(\mu, \sigma^2)$.

Equation (10) and the normality assumption allow us to derive the individual likelihood for any student i, given by

$$\begin{aligned} L_i(\Omega) =& f(\eta_{i,j,g,s,03-04}|\gamma^e, \theta, \rho, \mu, \sigma) \\ =& \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big\{ \frac{-1}{2\sigma^2} \cdot (y_{i,j,g,s,03-04} - y^C_{i,j,g,s,03-04} - \gamma_1^e e^N(\pi_{i,02-03}) \\ & - \theta e^N(\pi_{i,03-04}) - \rho \bar{e}^N_{s,02-03})^2 - \mu \Big\}. \end{aligned} \tag{11}$$

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

[36] This follows from the exclusive use of *school-level* growth targets under the ABCs, without corresponding student-level targets.

likelihood function:

$$\begin{aligned} l(\Omega) = & \sum_{i=1}^{N} lnL_i(\Omega) \\ = & -\frac{N}{2} ln(2\pi) - \frac{N}{2} ln(\sigma^2) \\ & - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \Big(y_{i,j,g,s,03-04} - y^C_{i,j,g,s,03-04} - \gamma^e_1 e^N(\pi_{i,02-03}) \\ & - \theta e^N(\pi_{i,03-04}) - \rho \bar{e}^N_{s,02-03} - \mu \Big)^2, \end{aligned} \tag{12}$$

where the full parameter vector is given by $\Omega = [\gamma^e_1, \theta, \rho, \mu, \sigma^2]'$.

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 3, justified above). When searching over values of γ^e_1 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^e_1.[37]

V.C.1 Identification

Separately identifying γ^e_1 and θ – the parameters related to NCLB – requires that, conditional on 2002-03 effort given by $e^N(\pi_{i,02-03})$, there is remaining variation in 2003-04 effort, denoted $e^N(\pi_{i,03-04})$, and vice-versa. This requirement is met in our application. Start with the non-monotonic shape of the effort function in 2002-03: this ensures that two students with the same level of effort in 2002-03 can have different levels of incentive strength under NCLB and, correspondingly, different levels of 2003-04 NCLB effort. Given the non-monotonicity of the 2002-03 effort function, the minimal condition for identifying the parameters is that the 2003-04 effort function should not be flat.

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, yet 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

[37] Specifically, on each iteration of the search, the routine selects a value for γ^e_1, substitutes that value along with the other known inputs into the known effort function $e^N(y^C_{i,j,g,03-04} + \gamma^e_1 e_{j(i,02-03)} - y^{T,N}_g)$ and generates an effort level in 2003-04 for each student. The iterative search continues until the routine arrives at a value for γ^e_1 that, together with the 2003-04 effort levels it implies, maximizes the log likelihood. In practice, we perform the estimation in MATLAB using the 'fmincon' command and supplying the gradient vector.

increases equally for each student, given 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.[38]

To identify the separate effects of NCLB and ABCs incentives and thus identify ρ, we draw on the fact that the ABCs incentives operate *across* schools while NCLB incentives operate *within* schools.[39] 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.

V.C.2 Maximum Likelihood Estimates

Having established parameter identification, we apply the maximum likelihood estimation routine to a sample of students observed in both 2002-03 and 2003-04 and for whom test scores are non-missing in both years. This restriction allows us to compare realized scores with counterfactual predicted scores for each student, needed to form the LHS of the main estimating equation.

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 carries forward 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. To rationalize this, it is likely that teachers and school principals realized over time that the sanctions might not be as binding as first thought.[40]

[38] 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. (As an aside, 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 function in 2003-04 that is not flat would be sufficient.)

[39] Specifically, 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.

[40] In practice, loopholes such as the Safe Harbor provision meant that sanctions were imposed less frequently

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 the following year.[41] While this is a relatively small effect, the positive and significant estimate of ρ is consistent with the effort response to NCLB in 2002-03 strengthening ABCs incentives in the following year by making the targets more difficult to pass, which in turn led to student performance gains. The estimates of γ^e_1 and θ are nearly identical to those obtained from the maximum likelihood routine that does not account for ABCs incentives, consistent with NCLB incentives varying *within* schools while ABCs incentives vary *across* schools (as the respective accountability incentives imply).[42]

In sum, our combined estimation approach has allowed us to recover the separate effects (current and persisting) of two unobserved teacher inputs – ability and effort – for the first time in the literature. The estimates for effort draw attention to the role of the economic agency of teachers. They make clear that teacher effort can be shifted, based on accountability incentive variation, and that the resulting effort brings performance benefits in the short and longer term.

VI Cost-Effectiveness Analysis: Effort versus Ability

We now explore policy implications of these new estimates. At a basic level, they suggest distinct input-specific policy levers for influencing student and school outcomes via shifts in ability and effort, respectively. Accordingly, we assess whether incentive reforms that target teacher effort are likely to be more cost-effective than policies that target teacher ability.

In order to do so, we present a novel framework that allows incentive reforms to be compared with alternative policies, based on their respective benefits (in terms of test score output) and associated financial costs. The framework has two key components: first, a method for credibly estimating the cost required to bring forth one extra unit of output under the incentive reform, and second, a means of cost-equating the two types of reform to ensure comparability. The resulting cost-effectiveness comparisons are new to the education policy literature.

than a strict enforcement of the rules would dictate.

[41] 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.

[42] The estimate of μ in the fullest specification in column (3) is 0.39 developmental scale points, or less than 5 percent of a student-level SD of the test score – not economically significant. It is also less precisely estimated than the other parameters.

We start this section by laying out the structure of the two policies we compare. Then we compute the benefits and costs of the ability-based reform, before describing the first and second key components of the framework. The headline numbers from the cost effectiveness analysis will make clear that the incentive reform comes out ahead by a substantial margin. We then present a thorough sensitivity analysis to explore the robustness of our findings, and also discuss possible extensions, given that our policy analysis represents a first step and not the last.

VI.A Two Types of Reform being Compared

Ability-based reforms seek to improve average teacher productivity through the dismissal of low-performing teachers. They have featured prominently in recent work, with Hanushek (2009, 2011), CFR (2014b), and Rothstein (2015) all analyzing policies that replace teachers whose value-added falls in the bottom portion of the measured distribution (for example, the bottom five percent).[43] Accordingly, we consider an *ability-based reform* that involves dismissing the bottom five percent of teachers in the VA distribution, drawing on that prior literature to provide credible estimates of the implied benefits and costs of the reform, useful for the comparison below.

The *incentive reform* we analyze uses the ABCs already in place in North Carolina as a template, which we scale to make it directly comparable to the ability-based reform. As with NCLB, it features performance targets, although these targets are student-specific rather than common to all students in a given grade. As such, *all* students are made marginal with respect to the scheme and not just a subset found close to a common threshold target. The incentive reform offers monetary rewards rather than non-pecuniary penalties (as under NCLB), which will allow a comparison to be made with the financial costs of the ability-based reform already estimated in the prior literature.

VI.B The Ability-Based Reform: Benefits and Costs

We now describe how the benefits (in terms of test scores) and the financial costs of the ability-based reform are calculated, drawing on the prior literature and our estimates above.

Benchmark estimates of the achievement benefits for ability-based reforms come from CFR (2014b). Those indicate that replacing teachers in the bottom five percent with random draws

[43]The literature has also focused on reducing the attrition of the highest rated teachers. Existing research (CFR 2014b) suggests that focusing on the top is less cost effective than replacing the lowest rated teachers, which motivates our focus on the latter.

from the distribution of new teachers results in an average two standard deviation improvement in teacher ability (measured in student test score units) for that subset. We take those values as a basis for estimating the benefits of a similar ability-based policy in our setting. A two standard-deviation improvement in performance for the five percent of teachers at the bottom of the distribution who are replaced translates into a performance improvement of 0.1 ($= 2 \times 0.05$) teacher-level standard-deviations across the full distribution of teachers – the short-run benefit of the ability-based reform.

To calculate the long-run benefit, we use our estimates of the persistence of teacher ability. These indicate that 19 percent of the initial effect persists four years into the future, implying that the ability-based reform achieves a performance improvement of 0.019 ($= 0.1 \times 0.19$) teacher-level standard deviations over that horizon. To express this long-run gain in terms of test scores, we note from Section IV.A that one (teacher-level) standard deviation is 1.79 developmental scale points, giving a long-run effect of the ability-based reform of 0.034 ($= 1.79 \times 0.019$) developmental scale points.

In terms of the costs, it is well-appreciated that ability-based reforms create 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 this additional cost of $700 per teacher.

VI.C The Incentive Reform

We wish to compare the ability-based reform with a viable incentive reform; here, the ongoing ABCs serves as a suitable template. In order to use that, we first need to recover the underlying mapping between output (measured by test scores) and expenditures (in dollars), the latter representing the pecuniary reward required to generate extra output through heightened effort incentives. We will think of this mapping in an incremental sense: the cost, based on the pecuniary rewards associated with ABCs-type accountability incentives, required to bring forth one additional unit of output. The incremental expenditure-output mapping in hand, we can then project the cost of attaining any given additional quantity of output. This will allow us,

via the second component of the framework, to calculate the benefits and costs of alternative settings of the incentive reform in ways that can be compared directly to other policies – in the current instance, the ability-based reform whose benefits and costs we have quantified.

The Output-Incentive Expenditure Mapping: We now explain how the output-expenditure mapping can be estimated in a credible way using the North Carolina data, starting with an intuitive account of our estimation procedure and introduce the key institutional features we draw on. We then lay out the series of steps involved in more detail, given that the approach is of independent interest.

Our primary goal (as noted) is to uncover the incremental dollar cost of a one-unit increase in output, measured by test scores, which can be written heuristically as dC/dy.[44] This response is not observed. It can, however, be recovered using the relationship:

$$\frac{dC}{dy} = \frac{dC}{de} \times \frac{de}{dy}, \tag{13}$$

noting that the two incremental responses on the RHS are estimable.

To estimate these components in our setting, we take advantage of two facts. First, the ABCs system uses monetary rewards, and second, there is necessarily a dynamic link between any NCLB effort response in 2002-03 (which is a response to non-pecuniary sanctions) and the ABC targets in 2003-04, given the way incentives are set under the value-added ABCs.

These facts allow us to leverage the introduction of NCLB in 2002-03 and the change in effort it engendered in that year (documented clearly above). Specifically, by altering the ABC's targets, the NCLB 'shock' provides exogenous variation in both financial rewards (the targets became harder to attain, lowering the expected payout) and also in effort, as schools responded to the induced ABCs incentive shock, leading to higher scores. These two sources of exogenous variation then allow us to estimate (i) the expected financial loss resulting from the increase in effort under NCLB (from which we can recover $\frac{dC}{de}$), and (ii) the test score effects of the increased effort response due to stronger incentives under the ABCs,[45] which can be inverted to give $\frac{de}{dy}$. We then combine these two effects, as in (13), yielding the desired mapping between expenditures and output that is central to the pecuniary reward-based class of schemes we focus

[44]The basic notation here is intended to fix ideas. We will introduce more precise notation when laying out the relevant steps in detail.

[45]Those incentives can be expressed as a function of school-level effort, as we will see.

on.

The Four Steps: We now describe the first component of the policy framework in more detail.[46] This involves the following four steps, also summarized in Table 5.

Step 1 - Calculating Schools' Expected Financial Losses Under the ABCs. School effort responses to the introduction of NCLB in 2002-03 raised the targets that schools faced in 2003-04 under the ABCs – automatically so, given the value-added incentives under the latter. Following the prescribed school growth score calculations under the ABCs and using the production technology given by equation (4), we estimate 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), writing this difference ΔF_s for school s,[47] where $F(\cdot)$ is the cdf of the school-level test score noise. Multiplying the differences in these passing probabilities by the ABCs bonus payment of \$750, received by teachers when their school satisfied its growth target, determines the expected per teacher dollar loss for each school because of its response to NCLB's introduction.

Based on our preferred estimate, the average school stood to lose \$122 per teacher in 2003-04 because of its effort response in 2002-03.[48]

Step 2 - Calculating the Change in Financial Incentives for a Unit Change in Effort. Next, to compute the change in 2003-04 financial incentives for a *one-unit* change in school effort from the previous year, we regress the expected dollar value each school stood to lose in 2003-04 (from Step 1) on average school-level effort from 2002-03. The resulting estimate $\hat{\beta}$ measures the response $\frac{d(750\cdot\Delta F_s)}{d\bar{e}^N_{s,02-03}}$. Intuitively, the underlying parameter governs the magnitude by which a one-unit increase in school-level effort in 2002-03, written $\bar{e}^N_{s,02-03}$, lowered the likelihood of ABCs

[46] Appendix F provides a thorough account.

[47] To be precise, we define ΔF_s, the difference in passing probabilities under the two scenarios, as

$$\Delta F_s = \Bigg[- F\Bigg(- \sum_{g=3}^{G_s} \left(\bar{y}^C_{g,s,03-04} - \alpha \bar{\tilde{y}}_{g-1,s,02-03}\right) - (\gamma^e_1 - \alpha)\bar{e}_{g-1,s,02-03} \Bigg) + F\Bigg(- \sum_{g=3}^{G_s} \left(\bar{y}^C_{g,s,03-04} - \alpha \bar{\tilde{y}}_{g-1,s,02-03}\right) \Bigg) \Bigg].$$

(See Appendix F for a detailed description of the notation involved and a full derivation.)

[48] This estimate in the number in bold font given in row (4) of Panel (a), Table 6. Other numbers reported in this subsection are drawn from the same column in that table, which we discuss in some detail below.

target attainment in 2003-04, expressed in terms of expected financial losses by multiplying by the ABCs bonus payment of $750.

In our central case, a one-unit increase in average school-level effort in 2002-03 would lead to an expected financial loss of $146.

Step 3 - NCLB's Effect on Subsequent Student Test Scores: Next, we need an estimate of the impact of lagged school-level effort on test scores in 2003-04. This is captured by the parameter ρ, which can be written explicitly as the incremental effect of school-level effort on scores, $\frac{dy}{d\bar{e}^N_{s,02-03}}$. It represents the effect of ABCs financial incentives (which NCLB alters – see Step 1) on test scores due to changes in teacher effort. We recovered this parameter precisely in Section V, estimating $\hat{\rho} = 0.29$.

Step 4 - Relationship between Test Scores and Changes in Financial Incentives: We then estimate the effect of test scores on financial incentives due to increased teacher effort by dividing the effect of lagged effort on subsequent financial incentives (from Step 2) by the effect of lagged school-level effort on subsequent test scores (from Step 3).[49]

Our preferred estimates indicate that the ABCs offering $504.70 per teacher would generate a one developmental scale point increase in test scores in the short run.

VI.D Policy Comparison Framework and Headline Estimates

The estimates of the underlying output-expenditure mapping feed into the second component of our policy analysis directly. This component allows us to place incentive reforms on a common footing with alternative policies, based on their respective costs and benefits.

In order to do so, we make the following scaling assumption:

Assumption 5 – Linear Scaling: a discrete change, ΔC, in the financial rewards associated with the candidate incentive reform will generate a discrete change in output, Δy, equal to $dy/dC \times \Delta C$.

[49] It is straightforward to see that dividing the effect of lagged school-level NCLB effort on expected ABCs financial rewards in 2003-2004 ($\hat{\beta} = \frac{d(750 \cdot \Delta F_s)}{d\bar{e}^N_{s,02-03}}$) by the effect of lagged school-level NCLB effort on test scores ($\hat{\rho} = \frac{dy}{d\bar{e}^N_{s,02-03}}$) 'cancels' the effort terms, giving the desired effect of test scores on financial incentives, $\frac{\hat{\beta}}{\hat{\rho}} = \frac{d(750 \cdot \Delta F_s)}{dy}$.

Under this assumption, the policy maker can compute the implied benefits and costs of the incentive reform in a way that is directly comparable to alternative policies (equating the financial costs and comparing the output benefits of each, for example) The cost (or benefit) projection that is implied utilizes the output-expenditure mapping estimated above; in an extensive robustness analysis that follows, we examine the sensitivity of the resulting projections.

To illustrate the approach, we focus on a comparison involving the ability-based reform described above. Recall that this reform costs an estimated $700 per teacher, and generates output (test score) benefits of 0.179 developmental scale points in the short run and 0.034 developmental scale points in the long run (after four years).

Our preferred estimates indicate a pecuniary incentive scheme with the same structure as the ABCs but offering $504.70 per teacher would generate a one developmental scale point increase in test scores in the short run. Using the linear scaling assumption to equate costs with the ability-based reform, the resulting test score increase produced by a cost-equating incentive reform that offered $700 per teacher would equal 1.39 (= $700/$504.70) developmental scale points in the short run.

To compute the long-run gain (in terms of test score 'output'), we impute the long-run persistence rate of effort by scaling our ML estimate of the one-year persistence rate (reported in Table 4). We do so in a way that gives the same rate of decay for teacher effort as the rate we estimate for teacher ability over a four-year horizon in Figure 5.[50] On that basis, the ABCs offering $700 per teacher would generate a long-run performance gain of 0.064 (or 1.39×0.0463) developmental scale points, greatly exceeding the corresponding benefit under the ability-based reform: for the same expenditure, the incentive reform is fully 88 percent more efficient.

Our policy comparison lends itself naturally to a simple graphical representation, in which each reform is shown as a point in output-expenditure (benefit-cost) space. This can be understood as the combination of the extra output and extra expenditure entailed by the given reform, relative to the status quo in which no policy is enacted.

Figure 8 illustrates the policy comparison we have in mind. The ability-based reform (represented by point A) has a known cost C_a and output y_a, provided by estimates from the prior literature. The incentive reform is represented by the linear cost function $C(y)$ with slope $\frac{dC}{dy}$,

[50]Specifically, Figure 5 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 the same pattern for the effects of effort, the effect four periods ahead amounts to 4.63% (i.e., 0.10×0.463) of the initial effort effect.

which we compute from $\frac{dC}{de}$ and $\frac{de}{dy}$ (by inverting $\frac{dy}{de}$), using the exogenous NCLB-based shifter of the ABCs target in 2003-04. The function makes clear the cost associated with any given output from the incentive reform. To compare the reforms, we select the output y_e that costs the same as the ability-based reform ($C(y_e) = C_a$). As drawn, the slope of the cost function implies that $y_e > y_a$, which means that the incentive reform is more cost effective than the ability-based reform. In general, the incentive reform is more cost effective if $\frac{dC}{dy} < \frac{C_a}{y_a}$; that is, the slope of the function is flatter than the line connecting the origin and point A.

VI.E Policy Analysis: Robustness and Extensions

The preceding framework permits the first cost effectiveness comparison to be made between incentive reforms and alternative education policies. The main estimates from applying the framework are striking. They indicate that implementing the incentive reform would involve a *substantial* benefit gain for the same cost as the ability-based reform under consideration – 88 percent more output for the same cost in our preferred specification – making them significantly more cost effective in the long run. This is despite the stronger estimated achievement effects for teacher ability than teacher effort. A central driver of the relative magnitudes of the new cost effectiveness estimates we provide is the following contrast: by design, ability-based reforms are only effective for a subset of teachers, while incentive reforms can be applied to all teachers.

Our policy framework allows the robustness of this cost effectiveness advantage to be explored in a systematic way. There are several relevant dimensions: (i) the variance of the school-level test score noise, a key determinant of the school passing probability under the ABCs accountability system; (ii) whether a short- or long-run perspective is taken; and (iii) the combination of parameters governing the incentive cost of an extra unit of output. We summarize the main findings along each dimension here.

Variance of the school-level test score noise. Table 6 shows the calculations that underpin the output-incentive expenditure mapping estimates based on our four-step procedure. The columns reflect different values of the standard deviation of the school-level test score noise. Our preferred estimate, already referenced, is a standard deviation of 0.36 developmental scale points.[51] Row (4) of the table shows that the implied expenditure under the incentive reform for one extra unit of short-run output is already less than the $700 per-teacher cost under the

[51] See Appendix F for the justification.

ability-based scheme (which delivers only 0.179 units of output) for all values of the SD other than SD=0.1; our preferred estimate yields a cost well below $700.

Time horizon. In Table 7, we compare the two types of reform over short- and long-run horizons. Over either horizon, the relative attractiveness of the incentive reform is increasing in the SD of the test score noise (shown across the table's columns). In the short run (one year), the incentive reform has a pronounced absolute advantage, reflected in the values in row (3) of Panel (a) – all above an output ratio of 4.5 for the same per-teacher cost of the reform. Because teacher effort is less persistent than ability, the advantage enjoyed by the incentive reform is less pronounced in the long run, especially for small values of the SD (see row (3) of Panel (b)).

Parameter values. Two key parameters determining the cost of generating one additional unit of output are the effect of effort on output (ρ) and the cost of higher effort (β), respectively. In Figure 9, we show combinations of the two parameters, below the upward-sloping line, for which the incentive reform is more cost effective in the long run. The point defined by our actual estimate for ρ – obtained from the ML procedure – and our preferred estimate for β falls well inside the region where the incentive reform has the cost advantage. The figure also depicts 95 percent confidence intervals around each parameter estimate, showing that almost the entire area defined by the confidence intervals of the two estimates falls within the region of ρ and β combinations where the incentive reform is more cost effective. As such, the incentive reform comes out ahead in a range of relevant cases.

The persistence rate of effort, γ_1^e, estimated to be 0.10 (see Table 4), is another important parameter relevant to the policy comparison. Increasing it has two effects. First, effort then persists more in the long-run (because we impute the four-year persistence rate from γ_1^e), implying the incentive reform achieves greater long-run output for the same cost. Second, increasing γ_1^e results in a smaller expected financial loss for each school under the ABCs (estimated in Step 1 above), lowering the cost of test score gains.[52] Both effects work to make the incentive reform relatively more cost-effective. Figure 10 captures this reasoning: it reproduces the main results from Figure 9, but shows how higher values of γ_1^e expand the region in which the incentive reform is more cost effective.[53]

[52]Greater student learning gains due to higher effort persistence make it less costly for schools to raise ABCs targets, resulting in a lower estimated value of β (from Step 2 above) – the financial cost of one more unit of effort, which also reduces the per-teacher cost of one more unit of output (given by $\frac{\beta}{\rho}$ in Step 4 above).

[53]Specifically, we consider the lower and upper bounds of the 95 percent confidence interval of our estimate,

In sum, the sensitivity analysis indicates that the incentive reform remains more cost effective than the comparable ability reform across a range of circumstances, in line with our main estimates.

Extensions: One can think of various ways in which the basic approach could be augmented: here, we list several. First, policies could be implemented that were a weighted average of the separate incentive and ability-based policies we compare. One might also consider the extensive margin effect of ability-based and/or incentive reforms on the *stock* of teachers, as the threat of dismissal or sharpened incentives could affect teacher turnover via participation constraints. Losses and gains may have asymmetric effects on teacher behavior, making reward-based incentive reforms more or less efficient than sanction-based reforms: it would be interesting to extend the analysis to allow for such potential asymmetry. In a similar vein, the use of extrinsic incentives could influence the intrinsic motivation of teachers, which would represent an additional cost of incentive-based policies.

Having noted these possible extensions, overall we view our policy analysis as a useful first step in placing incentive reforms in education on a comparable footing with alternative policies.

VII Conclusion

In this paper, we have shown how measured teacher performance is influenced by accountability incentives, in the process shedding light on the importance of teacher agency. Central to the analysis was an approach for separating out two unobserved education inputs for the first time: teacher effort, which is responsive to accountability incentives, from teacher ability, which is not. Our identification strategy leveraged a 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 incentive differences over time to separate teacher quality into teacher ability and the effort response to NCLB, allowing us to

equal to 0.06 and 0.14, respectively. (As above, we multiply these values by 0.463 to impute the four-year persistence rate, obtaining long-run estimates of 0.028 and 0.065.) As expected, raising γ_1^e to 0.14 rotates the black line up, resulting in a larger region of (ρ, β) combinations where the incentive reform has the advantage.

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 an approach built around the education production technology. This allowed us to identify the persistence of effort separately from teacher ability and 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. The estimates 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.

We then used the estimates and the technology to conduct a novel policy comparison. While incentive-focused education policies have become increasingly widespread over the past two decades, how they compare with alternatives has remained under-explored. This paper proposed an approach for computing the cost effectiveness of feasible policies (including incentive-based reforms) on a comparable basis for the first time in the education literature.

Our analysis indicated that using formal incentives constitutes a viable means of raising student and school performance. For the same per-teacher cost, we found that the incentive reform can deliver significantly higher output than comparable ability-based reforms – an advantage that holds across a variety of circumstances. This is attributable to the fact that incentive reforms can apply throughout the value-added distribution while ability-based reforms focus on a subset. Overall, the analysis shows that incentive reforms are worth considering seriously as a viable tool for education policy makers.

The general approach serves to open up a fuller comparison of the cost-effectiveness of alternative policies, based on refinements to the estimation strategies we develop – for instance, looking at the longer-run persistence of effort, and the teacher's explicit effort-setting decision itself. 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.

Chingos, Matthew M and Martin R West. 2012. "Do more effective teachers earn more outside the classroom?" *Education Finance and Policy*, 7 (1): 8–43.

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., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks. 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.

Fryer, Roland G., Steven D. Levitt, John List, and Sally Sadoff. 2012."Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment." NBER working paper 18237.

Fryer, Roland G. 2013. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." *Journal of Labor Economics*, 31: 373-407.

Goodman, Sarena F., and Lesley J. Turner. 2013. "The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program." *Journal of Labor Economics*, 31: 409-420.

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.

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

Jackson, Kirabo C., and Elias Bruegmann. 2009. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers." *American Economic Journal: Applied Economics*. 1(4): 85-108.

Jackson, Kirabo C. 2013. "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence From Teachers." *Review of Economics and Statistics*. 95: 1096-1116.

Jackson, Kirabo C., Jonah E. Rockoff, and Douglas O. Staiger. 2014. "Teacher Effects and Teacher-Related Policies." *Annual Review of Economics*. 6(1): 801-825.

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. 2011. "The Design of Performance Pay in Education" in Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, eds., *Handbook of the Economics of Education* vol. 4 (North-Holland: Amsterdam).

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.

Ost, Ben. 2014. "How Do Teachers Improve? The Relative Importance of Specific and General Human Capital." *American Economic Journal: Applied Economics*. 6(2): 127-151.

Papay, John P., Eric S. Taylor, John H. Tyler, and Mary Laski. 2016. "Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data." NBER Working Paper No. 21986. Cambridge, MA.

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.

Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2010. "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching," National Center on Performance Incentives.
http://www.performanceincentives.org/data/files/pages/POINT%20REPORT_9.21.10.pdf.

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	2,776,729
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[(b)]	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. They are used to measure the persistent effects of teacher ability and effort. We do not follow students past 2004-05, as the mathematics scale changes in 2005-2006, yet no table to convert scores back to the old scale was created by the state.
[(b)]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	-	-	-	-	-	-
Observations	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 performance-related teacher variables. 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 such distribution. The Observed SD is the raw standard deviation, while the Estimated SD 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 due to 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 Effort are calculated across all teacher observations in 2002-03.

Table 3 – The Effects of NCLB Incentives on Teacher Performance

	Third Grade		Fourth Grade		Fifth Grade	
	2002-03	Pre-NCLB	2002-03	Pre-NCLB	2002-03	Pre-NCLB
Effect of Incentives (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

Notes: This table presents estimates of the impact of incentives on teacher performance, where incentives are captured by the fraction of marginal students (m_{jt}) within classrooms and teacher performance is measured using teacher-year fixed effects. (The reported coefficients are given by 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 also includes year fixed effects. For third grade, the pre-NCLB years cover 1997 to 2000 and 2002, and for fourth and fifth grade, 1997 to 2001. 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 or 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)

Notes: This table presents maximum likelihood estimates of variants of equation (10). The sample includes fourth grade students in 2003-04. The dependent variable in each column is the difference between the realized and 'counterfactual predicted' mathematics score (see main text). The number of observations in each column is 86,236. Standard errors calculated using the Outer-Product of Gradients method appear in parentheses. *** denotes significance at the 1% level.

Table 5 – Recovering the Output-Incentive Expenditure Mapping – the Four Steps

Step	Description - What is Calculated	Formula[(a)]
Step 1	Expected financial loss under the ABCs, equal to the ABCs per-teacher bonus payment multiplied by the probability of passing the ABCs in 2003-04 (relative to the no-NCLB counterfactual scenario)	$750 \cdot \Delta F_s$
Step 2	Change in 2003-04 financial incentives for a *one-unit* change in school effort from the previous year	$\beta = \frac{d(750 \cdot \Delta F_s)}{d\bar{e}^N_{s,02-03}}$
Step 3	Impact of lagged school-level effort on test scores in 2003-04	$\rho = \frac{dy}{d\bar{e}^N_{s,02-03}}$
Step 4	Effect of test scores on ABCs financial incentives (due to increased teacher effort)	$\frac{\beta}{\rho} = \frac{d(750 \cdot \Delta F_s)}{dy}$

Notes: [(a)] The formulae in this column are referenced in the main text and derived in full in Appendix F. The term $\bar{e}^N_{s,02-03}$ in the denominator of Steps 2 and 3 measures school effort in 2002-03 in response to NCLB (relative to a baseline of zero).

Table 6 – Recovering the 'Output-Incentive Expenditure' Mapping – Sensitivity to the Variance of Test Score Noise

Standard Deviation of the School Error Term (ν_s)	0.1	0.2	0.3	**0.36**	0.4	0.5	0.6	0.7	0.8	0.9	1
Panel (a): Step 1 – Counterfactual School-Level ABCs Passing Probabilities and Expected Financial Loss											
(1) Average Passing Prob. if NCLB only operated in 2002-03	0.72	0.70	0.67	**0.66**	0.65	0.63	0.62	0.60	0.59	0.58	0.58
(2) Average Passing Prob. without NCLB	0.92	0.89	0.85	**0.82**	0.80	0.77	0.74	0.71	0.69	0.67	0.66
(3) Average Difference in Probabilities (ΔF_s)	-0.20	-0.19	-0.18	**-0.16**	-0.15	-0.14	-0.12	-0.11	-0.10	-0.09	-0.08
(4) Expected Per-Teacher Loss in Dollars ($C \equiv 750 \cdot \Delta F_s$)	-147.04	-141.81	-130.17	**-121.83**	-116.29	-103.16	-91.75	-82.13	-74.07	-67.30	-61.58
Panel (b): Step 2 – Change in 2003-04 ABCs Financial Incentives for a One-Unit Change in School Effort from Previous Year											
$\beta = \frac{dC}{de}$	-245.31 (26.97)	-199.45 (20.59)	-163.28 (16.27)	**-146.36 (14.42)**	-136.73 (13.42)	-116.62 (11.46)	-100.97 (10.02)	-88.60 (8.92)	-78.71 (8.03)	-70.68 (7.31)	-64.09 (6.70)
Panel (c): Step 3 – Impact of Lagged School-Level Effort on Test Scores in 2003-04											
$\rho = \frac{dy}{de}$ (same values in each column)				**0.29 (0.06)**							
Panel (d): Step 4 – Incentive Cost of One-Unit Increase in Output (Due to Increased Effort)											
$\frac{\beta}{\rho} = \frac{dC}{dy}$	845.91	687.75	563.03	**504.70**	471.47	402.14	348.18	305.53	271.40	243.73	220.98

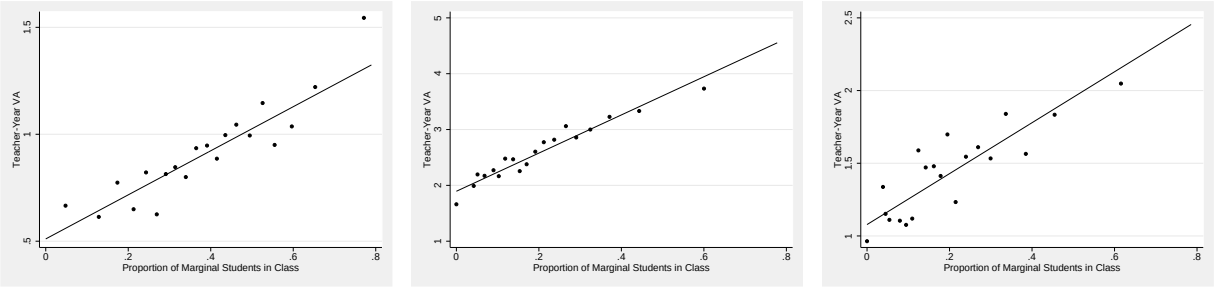
Notes: This table presents estimates for each of the four steps in the procedure to recover the output-expenditure relationship, described in Section VI.C. The unit of observation is a school in 2003-04 ($N = 1,250$). The columns correspond to different values of the standard deviation of the school-level error term (listed in the column headings); our preferred estimates are given in **bold**, under the "0.36" heading. In Panel (a), we calculate the expected per-teacher loss associated with the effort response to NCLB. Specifically, the average school passing probability is reported in row (1) for the counterfactual scenario in which NCLB only operated in 2002-03, and in row (2) for the counterfactual scenario in which NCLB was never enacted. Row (3) gives the average of the differences in passing probabilities for each school across the two scenarios. Row (4) multiplies this difference by the per-teacher bonus payment under the ABCs to calculate expected financial loss per teacher, averaged across all schools. In Panel (b), we regress the expected financial loss 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 parentheses. Panel (c) records the estimated effect of 2002-03 effort on test scores in 2003-04, common across all columns; it is the estimate of ρ from Table 4. Panel (d) reports the incentive cost of a one-unit increase in school-level output (due to increased teacher effort) as the SD of the school-level noise increases, from left to right.

Table 7 – Comparing Output under Incentive and Ability-Based Reforms
for the Same $700 Per-Teacher Cost - Sensitivity Analysis

Standard Deviation of the School Error Term (ν_s)	0.1	0.2	0.3	**0.36**	0.4	0.5	0.6	0.7	0.8	0.9	1
Panel (a): Short-Run Comparison											
(1) Incentive Reform Output (Dev. Scale Units)	0.827	1.017	1.243	**1.387**	1.484	1.740	2.010	2.291	2.579	2.872	3.167
(2) Ability Reform Output (Dev. Scale Units) – same values in each col.				**0.179**							
(3) SR Output Ratio: Incentive/Ability	4.623	5.686	6.945	**7.749**	8.294	9.724	11.23	12.799	14.409	16.045	17.696
Panel (b): Long-Run Comparison											
(1) Incentive Reform Output (Dev. Scale Units)	0.038	0.047	0.057	**0.064**	0.068	0.080	0.093	0.106	0.119	0.133	0.146
(2) Ability Reform Output (Dev. Scale Units) – same values in each col.				**0.034**							
(3) LR Output Ratio: Incentive/Ability	1.126	1.386	1.693	**1.882**	2.021	2.370	2.737	3.119	3.512	3.911	4.313

Notes: This table presents test score gains under cost-equated ability and incentive reforms in both the short run (Panel (a)) and long run (Panel (b)). In rows (1) and (2) of each panel, test score 'output' is measured in developmental scale units; for reference, the standard deviation of the student-level test score is 10 developmental scale units. In row (1) of Panel (a), output in each column under the incentive reform is determined as $700 \cdot dy/dC$, where dC/dy is calculated in Panel (d) of Table 6; in row (1) of Panel (b), output is calculated as $(700 \cdot dy/dC) \cdot \gamma_4^e$, where γ_4^e is estimated as 0.0463. Row (3) of each panel equals the ratio of row (1) to row (2) output. The columns correspond to different values of the standard deviation of the school-level error term (listed in the column headings); our preferred estimates are given in **bold**, under the "0.36" heading.

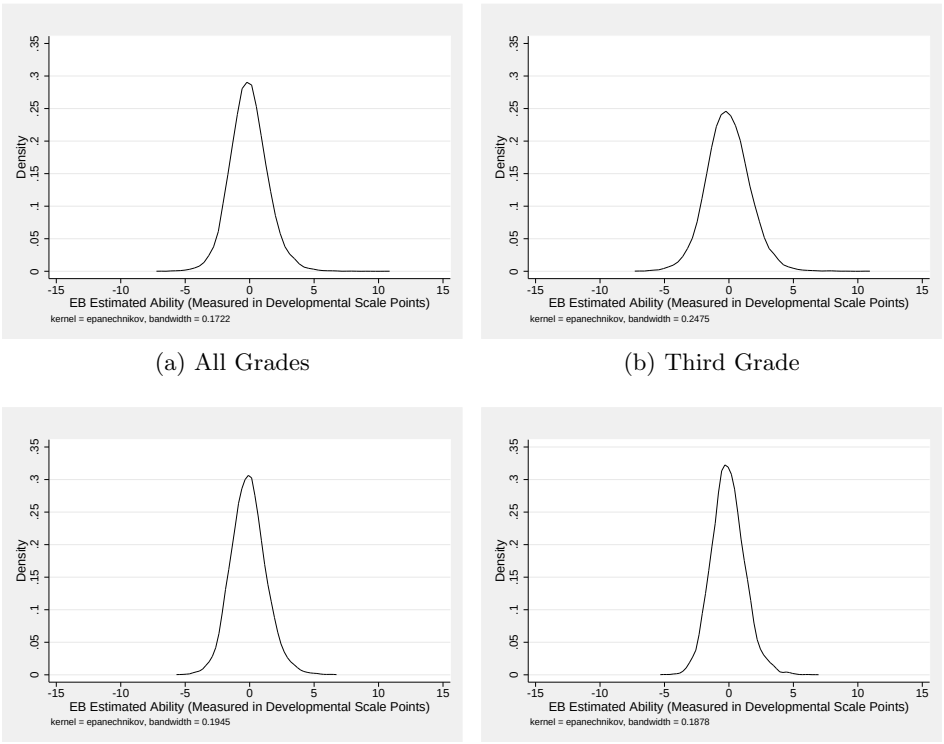
Figures



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

Notes: The panels in this figure plot 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 1 – Teacher-Year Fixed Effects versus the Proportion of Marginal Students in the Classroom

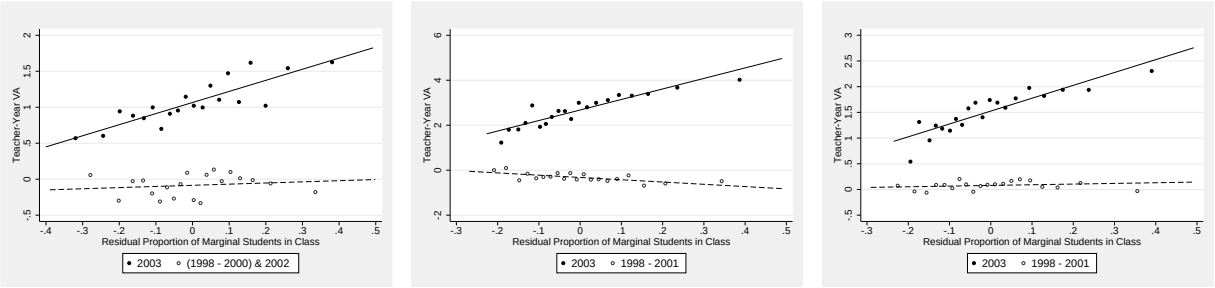


(a) All Grades (b) Third Grade

(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.

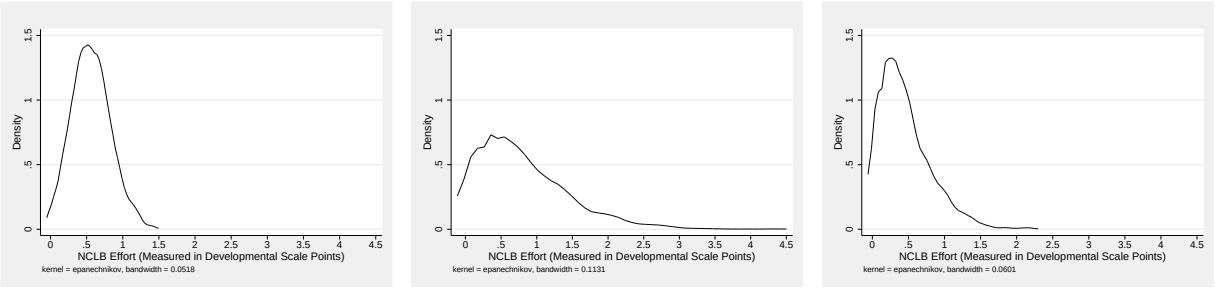
Figure 2 – Incentive-Invariant Ability Distributions



(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 are 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. 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 then calculate the average residualized m_{jt} and the average teacher-year effect. The circles in each panel represent these averages. The lines represent the estimated linear effects using the underlying teacher-year data.

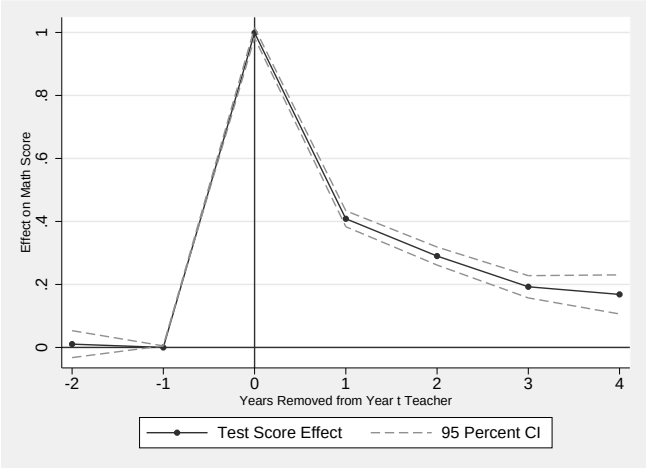
Figure 3 – Teacher-Year VA versus Proportion of Marginal Students in 2002-03 Classrooms



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

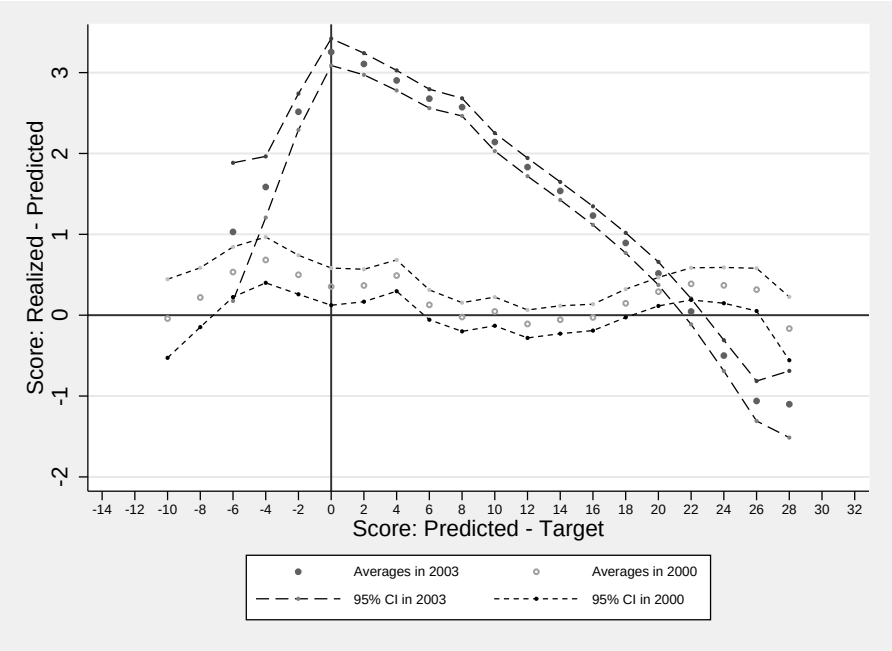
Notes: This figure illustrates teachers' effort responses to the introduction of NCLB. In panels (a) to (c), we present grade-specific densities of 2002-03 effort levels. To construct these figures, we first obtain 2002-03 effort for each teacher by taking the linear prediction (fitted value) from $e(m_{j2003})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

Figure 4 – Effort Distributions in 2002-03



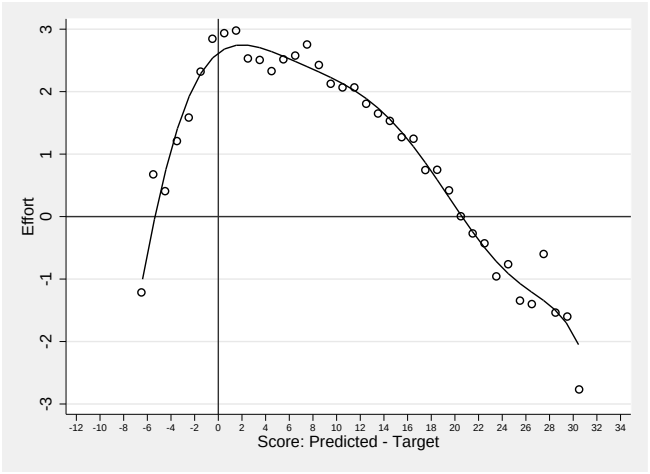
Notes: This figure reports estimates of the persistence of teacher ability (the ϕ_n coefficients) from equation (8). Each estimate is obtained from a separate regression in the pre-NCLB period from 1996-97 to 2001-02. The horizontal axis measures the number of years separating students from their period-t teacher while the vertical axis measures the impact of the period-t teacher on students' test scores in period $t + n$. The dots represent the estimated effects and the dashed lines, 95 percent confidence intervals with the associated standard errors clustered at the school level.

Figure 5 – Persistence of Teacher Ability and Baseline Effort in Pre-NCLB Period



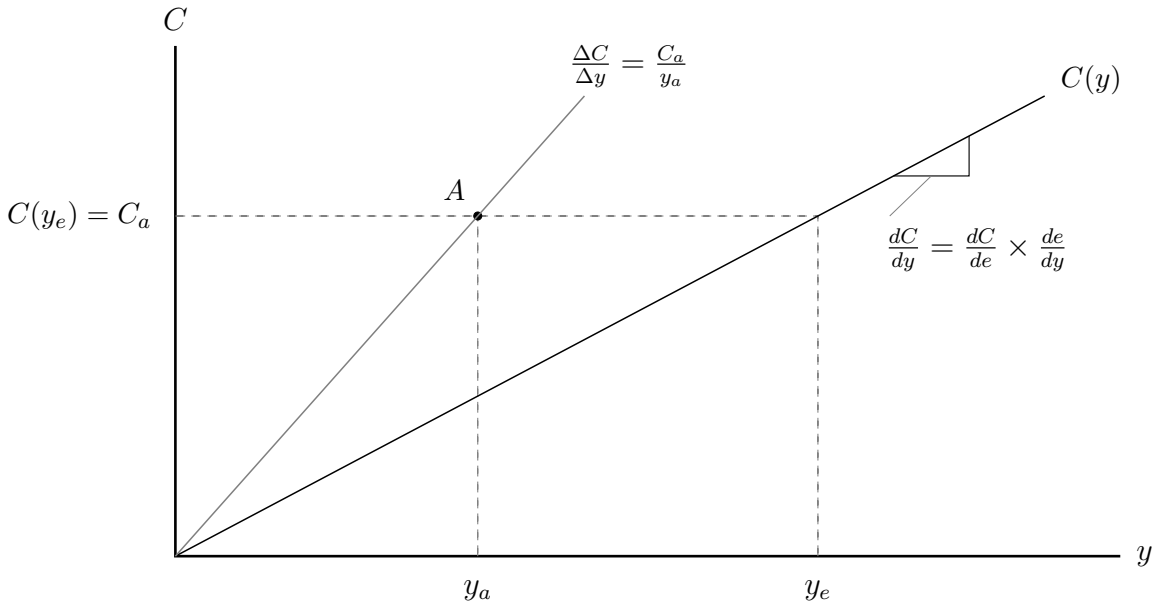
Notes: This figure shows the effect of accountability incentives on fourth grade mathematics scores. It is constructed as follows: In both years, we calculate a predicted score for each fourth grade student and then subtract off the known proficiency score target from this prediction – the horizontal 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 representing academic year 2002-03 averages, and the hollow circles, 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 6 – Inverted-U Response to NCLB's Introduction



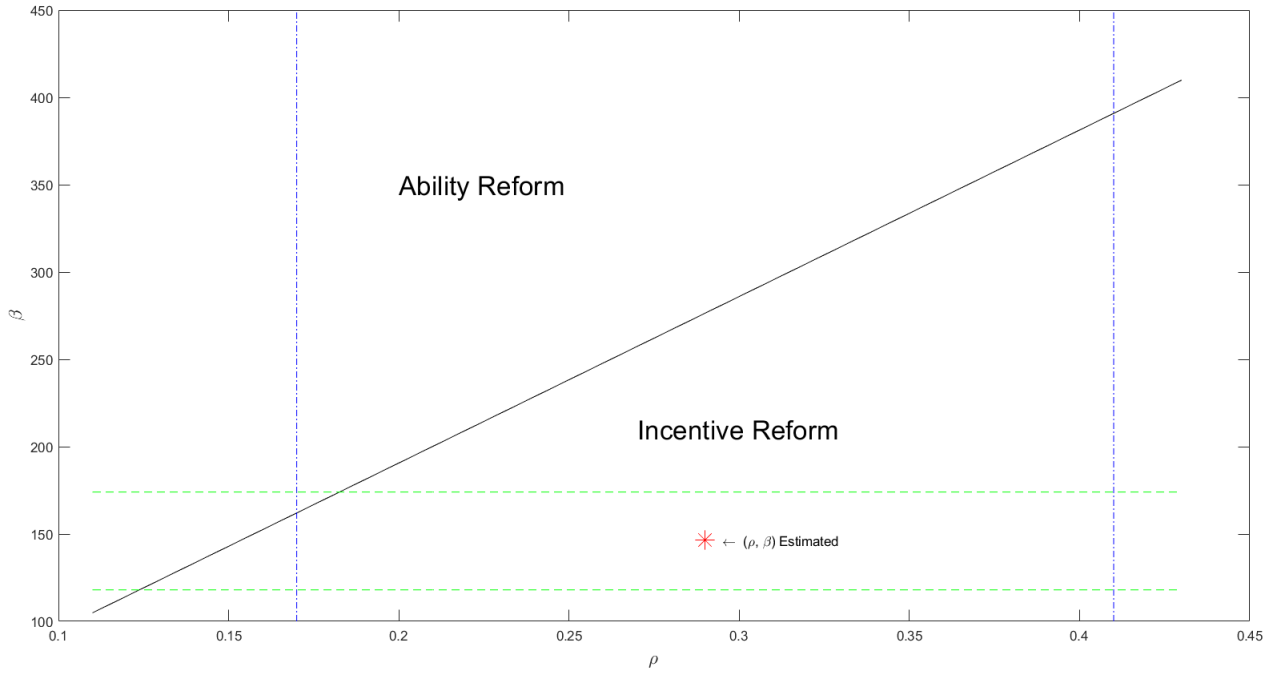
Notes: This figure plots the student-specific effort function. The horizontal axis measures incentive strength (as in Figure 6). To construct the smooth curve, we first form bins on the basis of incentive strength, then take the bin-specific differences between the year 2002-03 and the year 1999-00 vertical-axis variable in Figure 6. The circles 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 1999-00 and 2002-03) within each bin.

Figure 7 – Student-Specific Effort Function



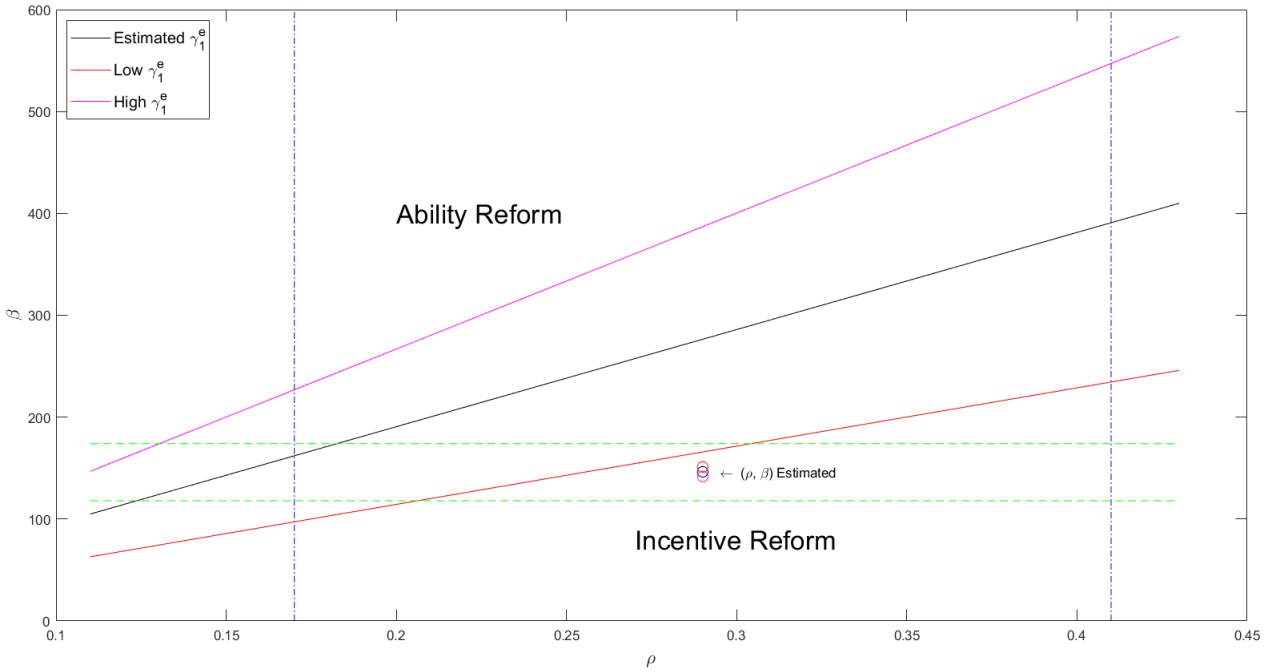
Notes: This figure illustrates the policy comparison we have in mind. The ability-based reform (represented by point A) has a known cost C_a and output y_a, provided by estimates from the prior literature. The incentive reform is represented by the linear cost function $C(y)$ with slope $\frac{dC}{dy}$, which we compute from $\frac{dC}{de}$ and $\frac{de}{dy}$ (inversion of $\frac{dy}{de}$), using the exogenous NCLB-based shifter of the ABCs target in 2004. The function makes clear the cost associated with any given output from the incentive reform. To compare the reforms, we select the output y_e that costs the same as the ability-based reform ($C(y_e) = C_a$). As drawn, the slope of the cost function implies that $y_e > y_a$, which means that the incentive reform is more cost effective than the ability-based reform. In general, the incentive reform is more cost effective iff $\frac{dC}{dy} < \frac{C_a}{y_a}$; that is, the slope of the function is flatter than the line connecting the origin and point A.

Figure 8 – Policy Comparison



Notes: This figure compares the long-run cost effectiveness of the ability and incentive reforms under various combinations of ρ and β. The incentive reform achieves higher output than the ability reform (for the same per-teacher cost of \$700) for all combinations of ρ and β below the solid black line, while the ability reform achieves higher output than the incentive reform above. The dashed vertical lines represent the 95% confidence interval for our estimate of ρ. The dashed horizontal lines represent the 95% confidence interval for β, based on a standard deviation of the school-level ABCs error term equal to our preferred estimate of 0.36 developmental scale points. The specific (ρ, β) estimates from the ML routine are indicated by the '*' in the figure.

Figure 9 – Comparing the Long-Run Cost Effectiveness of Ability and Incentive Reforms for Different (ρ, β) Combinations



Notes: This figure replicates Figure 9 under different assumptions for the value of the persistence rate of effort, γ_1^e. Refer to the notes of that figure for further details. Our estimated value, reported in Table 4 and used in the construction of Figure 9, is 0.10. The solid increasing black line from Figure 9 is reproduced here for ease of comparability, indicating the ρ and β pairs for which each type of reform is more cost effective under our baseline estimate of γ_1^e. In this figure, we additionally consider higher and lower values of γ_1^e, equal to 0.14 and 0.06 (the upper and lower bounds of the 95 percent confidence interval for our main estimate), respectively, and trace out how the cost-effectiveness region shifts in each case. The dashed vertical lines represent the 95% confidence interval for our estimate of ρ. The dashed horizontal lines represent the 95% confidence interval for β, based on a standard deviation of the school-level ABCs error term equal to our preferred estimate of 0.36 developmental scale points and using our baseline estimate for γ_1^e of 0.10. The specific (ρ, β) estimates from the ML routine are indicated by the 'o' in the figure. Changing the value of γ_1^e does not affect the estimate of ρ but does shift the estimate of β.

Figure 10 – Comparing the Long-Run Cost Effectiveness of Ability and Incentive Reforms for Different (ρ, β) Combinations
Under Different Assumptions for the Persistence of Effort (γ_1^e)

Appendices

Appendix A Relation to the Prior Literature

In this appendix, we place our analysis in the context of related prior studies, expanding on the discussion in the Introduction.

Among the studies estimating teacher value-added, most existing strategies in the literature treat teacher quality as fixed. Yet some recent evidence shows that teacher performance varies systematically over time and depending on the context – see Jackson, Rockoff, and Staiger (2014) for a detailed review. The workplace environment has been shown to matter particularly, as the interactions teachers have with their colleagues (Jackson and Bruegmann, 2009; Papay et al., 2016), the tasks-specific experience they accumulate (Ost, 2014), and the overall fit with their schools (Jackson, 2013) all affect measured performance.

Performance incentives – the central focus of our paper – also influence teacher quality. The existing literature has highlighted the contrasting effects of teacher pay-for-performance schemes on the one hand and school-level accountability programs on the other – see Neal (2011) and Figlio and Loeb (2011) for comprehensive reviews of each. Studies of teacher pay-for-performance schemes in the United States indicate that teacher-level programs are largely ineffective at improving teacher performance (Springer et al., 2010; Fryer *et al.*, 2012),[54] while the effectiveness of group-based programs depends heavily on program design (Goodman and Turner, 2013; Fryer, 2013; Imberman and Lovenheim, 2015). School-based accountability schemes, in contrast, have been found to change educators' actions in ways one would predict based on the incentives embedded in the programs. Proficiency-count systems, for example, have been shown to lead teachers to direct attention to marginal students, as expected (Reback, 2008; Neal and Schanzenbach, 2010; Deming et al, 2016), and growth-based programs can lead to resources being allocated across grades differentially in ways that make target attainment easier (Macartney, 2016).

Popular teacher VA estimators either seek to estimate *fixed* teacher quality (Kane and Staiger, 2008) or teacher quality that drifts over time according to a statistical process (CFR 2014a), despite the fact that school-based accountability programs are increasingly widespread in the US, and the evidence that these programs change teachers' actions. While these methods identify important heterogeneity across teachers, based on a fixed component of performance, they do not recover any information about the influence of (potentially changing) incentives on teacher performance.[55] As such, it is difficult to

[54] Fryer *et al.*, (2012) find large gains from a non-standard loss-aversion treatment but small gains from a traditional pay-for-performance program.

[55] For a given teacher, teacher VA methods typically construct some weighted average of the (residual) test scores of the students assigned to that teacher over time. If a teacher is observed many times and incentives are changing frequently, their influence in the teacher VA estimate may be averaged away. In contrast, when incentives remain constant over time, standard teacher VA estimates capture a composite of (true) fixed ability and the component of performance that is responsive to the incentive environment.

compare the effectiveness of policies that seek to change the distribution of the teaching force based on teacher VA with policies that aim to change the incentive environment in which teachers work.

We address these gaps by showing, first, that teacher VA estimates depend on both fixed teacher ability and the incentives teachers face from prevailing school-based accountability programs, thus highlighting the importance of the agency of teachers in determining overall teacher performance. We then compare the cost-effectiveness of incentive-based reforms alongside the widely-discussed ability-based reform that removes the lowest-performing teachers based on standard VA measures, showing that incentive reforms come out ahead in a range of plausible cases.

Our policy comparisons do not consider the potential for either reform to change the types of teacher who enter the teaching profession. Few existing studies address this issue in a North American context, although the limited available evidence suggests that both ability- and incentive-based reforms may induce positive selection into the teaching profession.[56] While teacher selection considerations are relevant when considering the generalizability of the effects of both types of reform, we note in our context that North Carolina implemented the state-operated ABCs of public education in the 1996-97 school year; this left ample time for selection to play out following the bonus payment reform prior to NCLB being implemented in the 2002-03 year. It is therefore unlikely that our proposed incentive-based reform would induce much additional selection into or out of teaching in North Carolina's public schools.

[56] Dee and Wyckoff (2015) find that a dismissal-based policy in the District of Columbia increased voluntary exit rates of low-performing teachers, suggesting a further gain from the ability-based reform associated with teacher self-selection. At the same time, financial incentives offered for high-performing teachers increased retention at the top of the performance distribution (although the effect was not statistically significant), indicating the performance incentives may operate in a similar manner. High-performing teachers may be induced to enter and stay in the profession given the opportunity to earn a bonus payment on top of regular income, as these teachers tend to have higher earnings opportunities outside the classroom (see Chingos and West, 2012).

Appendix B 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 in terms of our methods for – respectively – estimating teacher VA, predicting student test scores, and estimating the contemporaneous and persistent effects of teacher effort.

B.I Mathematics Scale Change

Mathematics scores were measured on different scales before and after 2000-01. Because North Carolina's ABCs accountability program required 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 that 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.

While test scores measured on both scales provide a valid way to track student learning, the timing of scale change and the steps involved in our analyses require us to use test scores expressed one scale or the other, depending on the task at hand. By way of overview, when reporting summary statistics, we keep a consistent presentation by expressing all mathematics 'level' and 'gain' scores on the first edition scale, except for third grade gains, which are calculated using first edition scores prior to 2000-01 and second edition scores thereafter. When estimating teacher VA, we also use test scores measured on the first edition in all cases except for third grade teachers observed after 2000-01 (whose VA we measure using the second edition). When estimating teacher effort, we only use test scores measured on the second edition scale because – as explained below – this allows us to predict students' test scores and their marginal status more accurately.[57] We now explain the implications of the test scale change for our analyses in more detail.

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

Teacher-Year Fixed Effects: When estimating teacher-year fixed effects for fourth and fifth grades, 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, based on equation (C.1) in Appendix C below. Here, both contemporaneous test scores (the dependent variable) and lagged test scores (control variables) are measured on the first edition scale, allowing us to measure fourth and fifth grade teacher VA

[57] It is worth noting that all of our results are qualitatively similar if we apply the test scale conversion and measure effort on the first edition, as we do for teacher value-added (results available on request).

on the first edition scale throughout the sample period. For third grade, because we are not able to convert the second grade test to the first edition scale, we conduct two separate regressions, given the prior-year test score is needed as a control variable in the value-added estimation: 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.

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 those teaching fourth and fifth grade. 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 for the pooled regression of students and teachers in fourth and fifth grades).

B.I.ii Scale Change Implications for Test Score Prediction

As explained in Section V of the main text and Appendix E below, a component 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.[58] 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.

[58] 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 to classify students correctly according to the distance between their predicted score and the proficiency cutoff, we conduct the analysis without converting.

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

As discussed in Section V of the main text and Appendix E 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 measured using the second edition when carrying out these exercises. Therefore, both initial NCLB effort and its persistence are estimated using mathematics test scores measured on the second edition scale.

B.II Reading Scale Change

We do not conduct our main analyses based on reading scores because the scale used to measure reading tests changed in 2002-03, coinciding exactly with the introduction of NCLB. As mentioned in Section IV of the main text, there is a potential concern that the timing of the scale change could have provided the state with an opportunity to change the curriculum or test scale in a way that would have allowed marginal students to perform better than in pre-NCLB years. 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 C Estimating Teacher Value-Added: Technical Details

This appendix describes the estimation of teacher VA in some detail: how the sample is selected for teacher VA estimation, how teacher-year fixed effects are estimated, and how we interpret Empirical Bayes (EB) estimates of teacher ability in light of North Carolina's pre-existing ABCs program.

C.I Construction of the Teacher Value-Added Sample

To estimate teacher value-added, 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 and year and if at least half of the tests administered by that teacher were for students in the correct grade.

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 existing studies using 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 reading 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.[59] Applying these restrictions leaves 1.67 million student-year observations for estimating teacher VA. Summary statistics for this sample are presented in Table C.1 below.

C.II Estimation of Teacher-Year Fixed Effects

As described in Section IV, we compute teacher-year fixed effects for each teacher using all students and teachers in the VA estimation sample. 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).[60]

[59] Special education and honors classes are excluded from the analysis, but students who repeat or skip grades are retained.

[60] Although standardizing test scores guards against changes in testing regimes over time, de-meaning would effectively remove the effects of changes in performance incentives: given our goal of assessing how teacher effort

Here we rely on the careful psychometric design of developmental scales in a North Carolina context, allowing one to track changes in learning within and across students as they progress through the education system.

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

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

Equation (C.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).[61] Conditional on those covariates, we obtain teacher-year fixed effect estimates as

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

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 IV. Summary statistics for our estimated teacher-year fixed effects are presented in Table C.2 below.

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

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

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

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

Accounting for the fact that the ABCs program was already operating in the pre-NCLB period, we

affects student learning, we wish to preserve all incentive-related performance variation over time.

[61] 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, which is likely to be correlated with innate student ability and previous teacher assignments.

can estimate the same equation but with modified notation, written

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

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-teacher comparison and should be unaffected 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[62] – a design feature that contrasts sharply with NCLB and ensures that effort incentives under the ABCs operate at the school level.[63] 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 (common 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 noted in the main text), given 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.

[62] The ABCs system 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.

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

Table C.1 – Student-Level Summary Statistics: 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[(b)]	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. They 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.
[(b)] The free or reduced-price lunch eligibility variable is not available prior to 1998-99.

Table C.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 D Robustness Checks

In this appendix, we consider the robustness of the results presented in Section IV.B. We also consider three plausible rival hypotheses to teacher effort setting: (i) differential sorting of students to teachers by ability, (ii) differential class size adjustments in response to NCLB, and (iii) differential sorting of students to classrooms based on peer characteristics.

D.I Alternative Definitions of 'Marginal Student'

First we demonstrate that the patterns in Figure 3 are robust to alternative cut-offs for defining a student as 'marginal.'

Figure D.1 below shows teacher VA in each grade as a function of the fraction of marginal students in the classroom in 2002-03 and placebo years (for many different definitions of 'marginal students'). Each panel of Figure D.1 shows an increasing relationship in 2002-03 and no relationship in the placebo years, lending credence to the claim that our results do not depend on the way we choose to classify students as marginal.

D.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 IV.B.1, we summarized several such hypotheses, which are explained in greater detail here.

D.II.i State and Family Responses

Curriculum Content or Test Design: One might worry that the state responded to the introduction of NCLB by designing new end-of-grade tests or by changing the curriculum in order to make it easier for marginal students to pass. In either case, marginal students would likely perform better than expected in 2002-03, possibly inducing the same positive relationship between the proportion marginal (m_{jt}) and teacher-year VA, independent of teacher effort. To address such concerns, we focus on the end-of-grade test in mathematics, as North Carolina used the same test and measured it on the same developmental scale before and after NCLB's introduction. The state also maintained the same achievement level test score thresholds, with Level Three corresponding to 'proficient' status both before and after 2002-03.[64]

[64] In contrast, the state issued a second edition of the reading test in 2002-03, the year in which NCLB took effect. Because of the coincident timing, we opt to not use the end-of-grade reading tests in our analysis.

Parental or Student Effort: Another concern is that marginal students and their parents reacted to the introduction of NCLB by adjusting their own effort. Such responses could generate an association between marginal student classroom presence and teacher-year VA, independent of teacher effort. Here we note that the introduction of NCLB did not create new stakes for students and parents, nor did it present them with new information. Thus it is unlikely that our strategy is affected by parental or student effort adjustments.

D.II.ii Other Potential School Responses

It is possible that schools reallocated students across classrooms in response to NCLB. For example, school principals might have assigned marginal students to higher ability teachers or smaller classrooms in 2002-03. In such cases, variation in marginal student presence across classrooms in 2002-03 would also reflect variation in other teacher and classroom characteristics that help determine teacher-year VA, calling into question whether our estimates reflect teacher effort. We address such concerns next, assessing whether differential sorting of marginal students to teachers based on (teacher) ability, class size, and several other classroom characteristics might explain the effects we estimate above.

Sorting Based on Teacher Ability: A natural way to gauge whether teacher sorting could be driving the results rather than additional effort being exerted by a *given* teacher is to test if 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 ($\hat{a_j}$), and an interaction between that term and an indicator for 2002-03:

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

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

Table D.1 shows the results from estimating variants of equation (D.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.[65] 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.[66] – for our main results to be biased upward, the

[65]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).

[66]To put the magnitude of the change in perspective, the estimate of β_2 in Table D.1 is 0.0033, implying that

change would have to be in the opposite direction.

While these analyses indicate that bias due to sorting by teacher ability is unlikely, we conduct a further test, showing that our main results are unchanged when estimating the effects of NCLB incentive strength using only *within-teacher* variation in performance, thereby removing any potentially confounding correlation between (fixed) teacher ability and incentive strength. To that end, we construct the difference between 2002-03 and 2001-02 teacher-year fixed effects for each teacher j as:

$$\begin{aligned}\hat{q}_{j02-03} - \hat{q}_{j01-02} &= a_j + e_{j02-03} + \bar{\epsilon}_{j02-03} - (a_j + e_{j01-02} + \bar{\epsilon}_{j01-02}) \\ &= e_{j02-03} - e_{j01-02} + \bar{\epsilon}_{j02-03} - \bar{\epsilon}_{j01-02}.\end{aligned} \tag{D.2}$$

The RHS of the first line is written in terms of ability and average effort components, and simplifies (on the second line) to differences in effort and noise.

We explore whether stronger NCLB incentives caused greater within-teacher performance improvements by relating the difference in teacher-year fixed effects to the fraction of marginal students faced by each teacher in 2002-03 ($m_{j,02-03}$). To account for mean reversion, we also control for a cubic function of 2001-02 teacher-year value-added, the estimating equation being

$$\hat{q}_{j02-03} - \hat{q}_{j01-02} = \alpha + \chi m_{j02-03} + g(\hat{q}_{j01-02}) + \zeta_{j02-03}, \tag{D.3}$$

where χ is the main parameter of interest, reflecting any relationship between NCLB incentives and within-teacher performance improvement, and $g(\hat{q}_{j01-02})$ is the cubic function of 2001-02 teacher-year VA.[67]

The panels of Figure D.2 show the partial relationships between the performance improvement in 2002-03 and $m_{j,02-03}$, while panel (a) of Table D.2 reports the underlying slope coefficients (i.e., estimates of χ), which are very similar to our main estimates in Table 3. Within-teacher performance improvements are clearly increasing in the fraction of marginal students in the classroom in 2002-03. Because the specification in equation (D.3) removes any effect of (fixed) teacher ability, it is unlikely that differential sorting of students to teachers based on ability can explain our results. A pooled regression of all pre-NCLB years (with transitions from year $t-1$ to t) is used as a placebo control in each grade, showing a relatively flat relationship, and further supporting the claim that the 2002-03 patterns reflect

a teacher who is one standard deviation (1.79 developmental scale points) below average had ($1.79 \times 0.0033 =$) 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.

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

NCLB effort incentives.

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 panel (b) of Table D.2, we investigate the importance of class size by including it as a control variable and replicating the analysis in Table 3. None of the point estimates are statistically distinguishable from their Table 3 counterparts.

Differential Sorting by Other Classroom Characteristics: As an additional robustness check, we account simultaneously for any potentially confounding effects from differential sorting of marginal students by teacher ability, class size, and a host of other classroom characteristics. We do so by pooling 2002-03 with all pre-NCLB years and estimating the following difference-in-differences regression:

$$\begin{aligned}\hat{q}_{jt} = \alpha_j + \gamma_t + \psi_1 m_{jt} + \psi_2 1(t = \text{2002-03}) * m_{j02-03} \\ + w(exp_{jt}) + \omega X_{jt} + \xi_{jt}, \forall\ t \leq \text{2002-03}.\end{aligned} \tag{D.4}$$

This involves regressing teacher-year fixed effects on teacher fixed effects, year fixed effects, the proportion of marginal students in the classroom, this proportion interacted with a NCLB-period indicator, controls for teacher experience, and several other classroom characteristics.[68]

The main parameter of interest is ψ_2, which captures the differential relationship between teacher-year performance and the fraction of marginal students in the year NCLB is introduced, relative to the relationship that prevailed in the pre-reform period. The inclusion of teacher fixed effects and classroom characteristics in equation (D.4) guarantees that ψ_2 is identified using within-teacher variation, conditional on many classroom characteristics. If NCLB incentives operate independently of teacher ability and other classroom characteristics, we would expect to find a positive and significant estimate for ψ_2.

The results from estimating variants of equation (D.4) are reported in Table D.3.[69] In column (1), we do not include teacher fixed effects or classroom characteristics, instead accounting for teacher ability using our EB estimates (which, again, are jack-knife or leave-year-out

[68]Specifically, the vector of classroom characteristics (X_{jt}) includes class size, the classroom average prior math score, the classroom average prior reading score, and the fractions of students who are racial minorities, disabled, limited English proficient, male, and who have college-educated parents.

[69]The table reports estimates for the sample of fourth grade teachers. The results in third and fifth grade are similar.

estimates in pre-NCLB years). The results are very similar to our main results above, as the relationship between the fraction of marginal students and teacher performance is large and positive in 2002-03 but not in the pre-reform period. Accounting for teacher fixed effects in column (2) yields very similar estimates, as does dropping the teacher fixed effects while accounting for classroom characteristics in column (3). In column (4), we account for both teacher fixed effects and classroom characteristics and again obtain nearly identical results, suggesting that differential sorting of marginal students to teachers based on ability or to classrooms based on various characteristics is not first order.

In sum, the evidence buttresses our effort interpretation – none of the robustness analyses supports the view that differential sorting is driving our main results. Further, we note that our estimates of the effects of NCLB incentives on teacher performance are not dependent on the method we use to estimate teacher ability. That is, the specifications given by equations (D.3) and (D.4) are agnostic as to how teacher incentive-invariant ability is measured, as they do not require an estimate of ability, yet produce very similar estimates to our main specifications (where teacher ability is estimated using the EB procedure).

Table D.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 (D.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 EB estimate in pre-NCLB years to avoid any mechanical correlation between EB estimates and outcomes. Standard errors clustered at the school-level appear in parentheses. *** denotes significance at the 1% level, and ** denotes significance at the 5% level.

Table D.2 – Robustness Checks: Sorting by Teacher Ability and Class Size

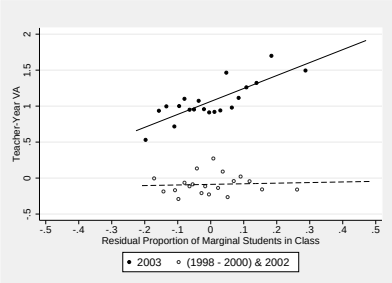
Panel (a): Change in Teacher-Year VA as Dependent Variable						
	Third Grade		Fourth Grade		Fifth Grade	
	2002-03	Pre-NCLB	2002-03	Pre-NCLB	2002-03	Pre-NCLB
Effect of Incentives (m_{jt})	2.08***	0.18	4.11***	-0.16	2.48***	0.49***
	(0.18)	(0.16)	(0.31)	(0.18)	(0.24)	(0.17)
Observations	2,651	9,697	2,453	9,087	2,397	8,357
Panel (b): Teacher-Year VA as Dependent Variable Controling for Class Size						
	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: In Panel (a), we regress the change in teacher-year VA on incentive strength (m_{jt}) in the year NCLB was introduced (2002-03) and in pre-NCLB years. (The coefficient estimates are for χ from grade-specific regressions of equation (D.3).) Specifically, for 2002-03, 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. In the pre-NCLB period, we regress the change in teacher-year VA from year $t-1$ to t (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.
In Panel (b), we present estimates of the effects of the fraction of marginal students within classrooms on teacher performance. (The reported coefficients are 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 estimates in the pre-NCLB columns come from pooled regressions of all pre-NCLB years that also includes year fixed effects. For third grade, the pre-NCLB years cover 1998 to 2000 and 2002, and for fourth and fifth grade, 1997 to 2001. Standard errors clustered at the school level appear in parentheses. *** denotes significance at the 1% level.

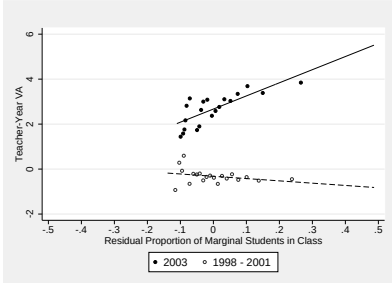
Table D.3 – Difference-in-Differences Estimates

	(1)	(2)	(3)	(4)
Incentive Strength (m_{jt})	-0.88***	-0.17	-0.44**	-0.23
	(0.15)	(0.23)	(0.19)	(0.25)
Incentive Strength post-NCLB ($m_{jt} \times (t = 2002/03)$)	5.45***	6.21***	5.32***	5.91***
	(0.35)	(0.54)	(0.36)	(0.55)
Teacher Fixed Effects?	N	Y	N	Y
Classroom Characteristics?	N	N	Y	Y

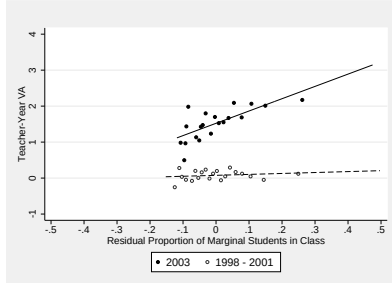
Notes: In this table, we present difference-in-differences estimates from regressions of teacher performance on incentive strength and incentive strength interacted with a post-NCLB indicator. (The estimates correspond to coefficients ψ_1 and ψ_2 from equation (D.4).) The dependent variable in each specification is a teacher-year effect. Incentive strength is given by the fraction of marginal students in the classroom. All regressions include indicator variables for teacher experience (in bins). Specifications without teacher fixed effects control for incentive-invariant teacher ability using EB measures of teacher ability. Classroom characteristics include class size, the classroom average prior mathematics score, the classroom average prior reading score, and the fractions of minority, disabled, limited English proficient, male, and students with college-educated parents. The sample consists of fourth grade teachers between 1997-98 and 2002-03; results for third and fifth grade are qualitatively similar. The number of observations in each column is 14,149. Standard errors clustered at the school-level are given in parentheses. *** denotes significance at the 1% level, and ** denotes significance at the 5% level.



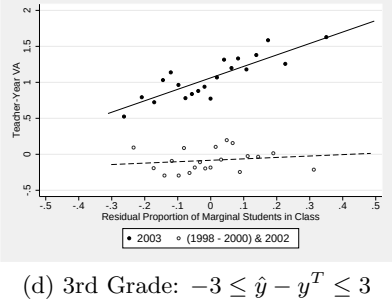
(a) 3rd Grade: $-2 \leq \hat{y} - y^T \leq 2$



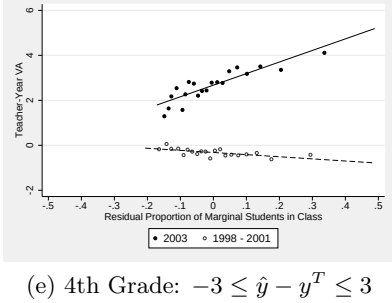
(b) 4th Grade: $-2 \leq \hat{y} - y^T \leq 2$



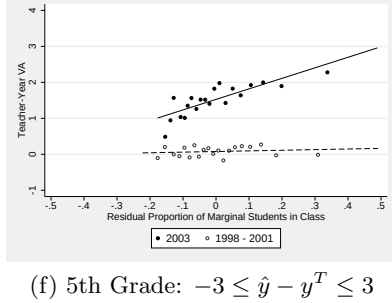
(c) 5th Grade: $-2 \leq \hat{y} - y^T \leq 2$



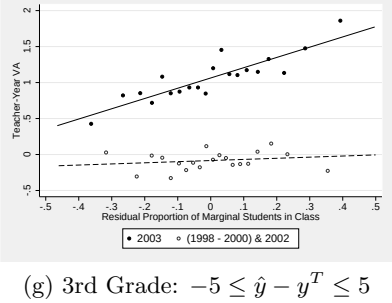
(d) 3rd Grade: $-3 \leq \hat{y} - y^T \leq 3$



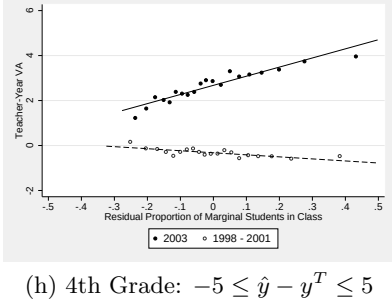
(e) 4th Grade: $-3 \leq \hat{y} - y^T \leq 3$



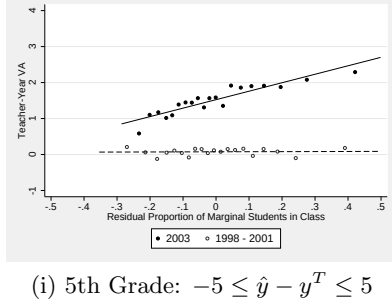
(f) 5th Grade: $-3 \leq \hat{y} - y^T \leq 3$



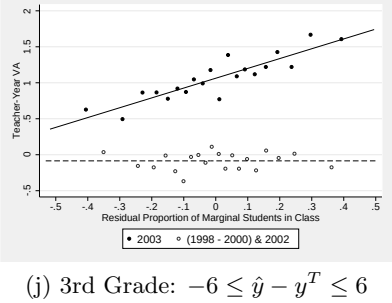
(g) 3rd Grade: $-5 \leq \hat{y} - y^T \leq 5$



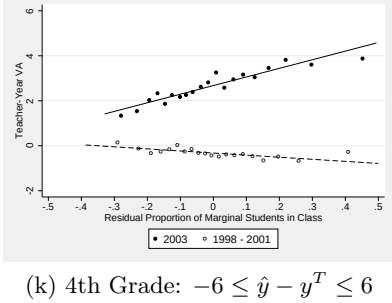
(h) 4th Grade: $-5 \leq \hat{y} - y^T \leq 5$



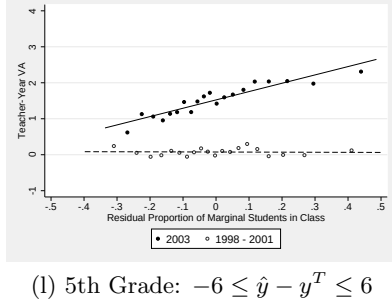
(i) 5th Grade: $-5 \leq \hat{y} - y^T \leq 5$



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



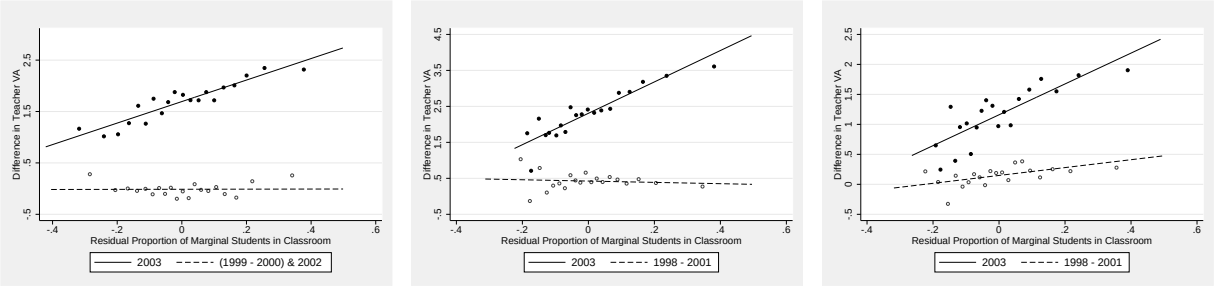
(k) 4th Grade: $-6 \leq \hat{y} - y^T \leq 6$



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

Notes: This figure reproduces the analysis in Figure 3 with alternative cutoffs for marginal student status. See the notes of Figure 3 for details. The range for marginal student classification is given in the label of each panel.

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



(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 (D.3). 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 D.2 – Within-Teacher Performance Improvements

Appendix E Estimating the Persistence of Effort: Technical Details

This appendix presents our methodology for estimating the persistence of teacher effort, supplementing the discussion in the main text. We first summarize the key variables used in our approach, then describe how teacher effort is estimated at the student level, before deriving the main estimating equation used in our maximum likelihood procedure.

E.I Definitions

The table below presents the definitions of the key variables used to exposit our approach, along with the corresponding notation..

Table E.1 – Estimating the Persistence of Effort – Notation and Definitions

Variable	Definition
$\hat{y}_{i,j,g,s,02-03}$	The predicted mathematics score of student i, assigned to teacher j in grade g at school s in academic year 2002-03.
$y^{T,N}_{g,t}$	The mathematics test score proficiency target in grade g mandated by NCLB in year t. (The superscripts T and N indicate 'target' and 'NCLB,' respectively.)
$\pi_{i,02-03}$ $(\equiv \hat{y}_{i,j,g,s,02-03} - y^{T,N}_{g,02-03})$	Incentive strength for student i 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, assigned to teacher j in grade g at school s in academic year 2003-04, under the scenario where counterfactual NCLB effort is zero.
$\pi_{i,03-04}$ $(\equiv y^{C}_{i,j,g,s,03-04}+\gamma^{e}_{1}e^{N}(\pi_{i,02-03})-y^{T,N}_{g,03-04})$	Incentive strength for student i 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, as shown in Figure 7. In academic year 2002-03, this function takes $\pi_{i,02-03}$ as its argument: in 2003-04, it takes $\pi_{i,03-04}$ as its argument.
γ^{e}_{1}	The one-period forward persistence rate of NCLB effort.

E.II Estimating Student-Level Effort

To construct a measure of effort at the student level, we draw on the semi-parametric patterns in Figure 6 (described in the main text). These show that the introduction of NCLB had clear

non-linear effects on student test scores, consistent with strong teacher effort responses to the incentive scheme. This interpretation draws on two notions: (i) a student's *predicted score*, and (ii) a student-specific measure of *incentive strength*. We discuss each in turn, before explaining how they are used to derive the student-level effort measure we use.

E.II.i Predicted Score

The predicted score for each student in 2002-03, which captures the score the student would have earned in that year had NCLB not been enacted, is calculated in two steps. First, we predict student performance in a flexible way regressing contemporaneous 2001-02 mathematics scores on various controls in the pre-NCLB period,[70] and save the estimated coefficients. Second, we make an out-of-sample mathematics test score prediction for students in 2002-03 by combining the estimated coefficients from the first step with the (pre-determined) covariates of students in 2002-03, denoting the predicted score for student i in 2002-03 by $\hat{y}_{i,j,g,s,02-03}$.

Next, we use equation (3) to express this predicted score in terms of parameters of the technology – key to being able to uncover NCLB-related effort in 2002-03.

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

The RHS of the prediction consists of only non-effort inputs, setting effort to zero along with any prediction error.[71] Intuitively, this prediction represents the score students would have earned had teachers not adjusted their effort decisions in response to NCLB's introduction, given it is estimated using the relationship between student characteristics and test scores that prevailed prior to NCLB.

We then use student i's predicted score together with the student's realized test score in 2002-03, $y_{i,j,g,s,02-03}$, to obtain an estimate of the effort (plus noise) received by each student. In

[70] These include 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.

[71] Although the prediction error need not be zero at the individual level, our approach for estimating NCLB effort requires only that it be zero *on average*, given that we rely on mean differences between realized and predicted scores throughout the predicted score distribution. We demonstrate below that these mean differences are indeed centered around zero prior to NCLB's enactment, supporting the mean-zero assumption.

particular, taking the difference between the realized and predicted scores for student i yields:

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

This differencing gives the sum of NCLB effort and a random shock to test scores on the last line.[72]

E.II.ii Incentive Strength

Our measure of *incentive strength* is given by the distance between the predicted score (just described) and the fixed NCLB proficiency target. Letting $y_{g,t}^{T,N}$ denote this NCLB target in grade g in year t (where 'T' in the superscript stands for 'target' and 'N' stands for 'NCLB'), we define incentive strength in 2002-03 as:

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

This measure is used on the horizontal axis in Figure 6, 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 (E.1)) within each bin, recovering average teacher effort as a function of incentive strength under the assumption that the idiosyncratic test score noise has a mean of zero.

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

Figure 6 makes clear that students predicted to score near the proficiency threshold – those for whom effort incentives are strongest – do on average receive the biggest boost to their scores. To ensure that we do not systematically under- or over-predict for certain parts of the test score 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 well throughout the distribution, given by the approximately flat line; this supports the view that the 2002-03 patterns reflect student-specific NCLB effort.

[72] The vertical axis in Figure 6 (discussed below) plots averages of the differences given by equation (E.1), eliminating the influence of test score noise.

We then use the profiles for the two years in Figure 6 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, where the weights are given by 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 7. We use this function to assign a level of effort to each student directly, according to the following assumption:

Assumption E.1: Given student-specific values of $\pi_{i,02-03}$ and the estimated effort function $e^N(\cdot)$, the effort directed by teacher j to each student i in 2002-03 is given by $e_{j(i,02-03)} = e^N(\pi_{i,02-03})$.

E.III The Components of the Estimating Equation

With the student-specific effort measure in hand, we turn to specifying an equation for estimating effort persistence and other relevant parameters that can be taken to the data.

We start by using (4) to obtain an expression for test scores in 2003-04 as a function of inputs in that year and inputs from the previous year whose effects persist:

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

The RHS of this equation captures, on the first line, the persistent effect of once-lagged scores from 2002-03 excluding teacher ability or effort, written $y_{i,j',g-1,02-03} - a_{j(i,02-03)} - e_{j(i,02-03)}$, where the j-subscripts coincide – that is, $j' = j(i, 02-03)$. Given the technology, we interpret this component as the persistent effect of non-ability and non-effort inputs. The second line captures the persistent effects of teacher ability from 2002-03 and teacher ability in the current year 2003-04; and the third line includes the persistent effects of teacher effort from 2002-03 and in the current year 2003-04, along with a random shock to current test scores.

From an estimation perspective, the first two lines of the RHS consist of components that can all be estimated using the prior steps of our estimation procedure. On the third line, the only

component that remains to be estimated is contemporaneous effort, $e_{j(i,03-04)}$, noting the first term, $e_{j(i,02-03)}$, can be recovered from the procedure expressed in Assumption E.1. Effort in 2003-04 needs to be controlled for in the estimation procedure to avoid overstating the persistent effect of effort from the previous year. Given the prevailing incentives in 2003-04, we hypothesize that contemporaneous effort comes from two sources: NCLB, introduced in the previous year; and the ABCs, whose incentives are likely to have been disrupted following the effort response to NCLB in 2002-03. We take these two sub-components in turn.

E.III.i NCLB Effort in 2003-04

Under NCLB, the descriptive evidence indicates that marginal students are likely to receive the most teacher effort. To identify marginal students in 2003-04, and thus who is likely to receive more effort and who less, we need to form a test score prediction in the absence of NCLB effort being exerted in that year. This will determine who, absent such effort, is likely to be close to the 2003-04 target and who will be further away.

Here, it is convenient to introduce some notation that will both help to form that test score prediction and to express the main estimating equation in a convenient way.

Definition E.3 – Counterfactual Predicted Score in 2003-04: $y^{C}_{i,j,g,s,03-04} \equiv \gamma(y_{i,j',g-1,s',02-03} - a_{j(i,02-03)} - e_{j(i,02-03)}) + a_{j(i,03-04)} + \gamma^{a}_{1} a_{j(i,02-03)}$.

This notation collects the first two lines of (E.2) under the label $y^{C}_{i,j,g,s,03-04}$. We think of this as the test score that students *would* have earned in 2003-04 had NCLB not been enacted in the prior year.[73] The term "counterfactual" reflects the notion that, having isolated the inputs other than effort that determine test scores in 2002-03, the scores that would prevail *counterfactually* in 2003-04 would solely reflect the persistent effects of non-effort inputs from the previous year plus contemporaneous (2003-04) ability. In effect, we envisage a counterfactual effort decision on the part of teachers that involves setting 'counterfactual' effort in both years equal to zero – notationally, $e^{C}_{j(i,02-03)} = e^{C}_{j(i,03-04)} = 0$.[74]

[73] To be very clear, NCLB *did* occur in that prior year, and did engender an actual effort response, denoted by $e_{j(i,02-03)}$. Thus it is necessary to subtract actual effort from test scores in 2002-03 in order to isolate non-effort (and non-ability) inputs, whose persistent effects we wish to keep track of in 2003-04; the technology assumptions imply the appropriate 'recipe' for purging effort (and ability) effects from once-lagged scores.

[74] In the counterfactual scenario, simply plug zero counterfactual effort levels on the last line of (E.2), and this yields $y^{C}_{i,j,g,s,03-04}$, according to the definition (assuming the test score noise is zero in expectation).

In our estimation procedure, we will use an empirical analog to the counterfactual predicted score in 2003-04. This is constructed along the same lines as the predicted score ($\hat{y}_{i,j,g,s,02-03}$) in the previous year, which also abstracted from NCLB effort. Specifically, we use the prediction equations (one for each grade) from the first step of the procedure for estimating student-level effort above to make forecasts for 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 $g-1$ test scores, forecasting based on the actual 2003-04 values for all other covariates. Realized grade $g-1$ scores from 2002-03 contain NCLB effort, and their persistence into 2003-04 therefore depends on the effort persistence parameter, 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 in the absence of NCLB incentives in 2002-03.

Using the machinery just developed, we now discuss teacher effort setting in response to NCLB in 2003-04. Here, additional structure is needed to capture the way teachers form predictions (drawing on the notation we introduced) about likely student performance in 2003-04 and specifically, how they incorporate the persistence of prior-year effort into those predictions. We make the following pair of assumptions – about teachers' information sets and the common test score prediction rule they follow, respectively:

Assumption E.2a: 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 E.2b: The prediction teachers make about each student's contemporaneous test score (in 2003-04) in the absence of any contemporaneous effort is given by $y^C_{i,j,g,s,03-04} + \gamma_1^e e_{j(i,02-03)}$.

The prediction draws on the technology directly. According to equation (E.2) and using the definition of y^C above, student test scores in 2003-04 can be written:

$$y_{i,j,g,s,03-04} = y^C_{i,j,g,s,03-04} + \gamma_1^e e_{j(i,02-03)} + e_{j(i,03-04)} + \nu_{i,j,g,s,03-04}\,,$$

which is the predicted score in that year plus contemporaneous teacher effort (plus noise). In line with the previous assumption, teachers use the student-level prediction in 2003-04 to decide how much effort to devote to each student, taking account of the incentive in the current year

to direct effort to student i. That incentive is given by

Definition E.4 – Incentive Strength in 2003-04: $\pi_{i,03-04} \equiv y^{C}_{i,j,g,s,03-04} + \gamma^{e}_{1} e_{j(i,02-03)} - y^{T,N}_{g,03-04}$.

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 an effort-setting rule expressed in

Assumption E.3: 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 rely on 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 assumption, 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 NCLB incentives remained in place across 2002-03 and 2003-04.[75] Yet 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 reflecting changes in effort over time.[76]

Given Definition E.4 and Assumption E.3, 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^{e}_{1}, as these influence incentive strength in 2002-03, highlighting the correlation of effort over time. We will account for this correlation in our main estimating equation.

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

[76] As an alternative to having effort levels changing proportionally across years, one might think that teachers would become better able to predict which students were marginal over time, thereby directing more effort to those students, which would result in a compressed effort function rather than one that is scaled up or down by a constant factor. This alternative hypothesis is unlikely to hold 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 system 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. 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.

E.III.ii Accounting for ABCs Effort

The other subcomponent of 2003-04 effort referenced above is associated with changes to effort incentives under North Carolina's pre-existing ABCs program in 2003-04. Such changes arise because of teacher effort responses in 2002-03 following NCLB's introduction. ABCs incentives will be affected (in ways we describe below) by changes in prior scores, given that ABCs growth targets depend on those scores by institutional design.

Our strategy to account for changing ABCs effort decisions draws on institutional features of the ABCs system and our conceptual framework. In terms of relevant institutional background, North Carolina's ABCs program sets test score *growth* targets that are grade- and subject-specific for each school. The targets for average test score growth across all students in a subject-grade are a linear function of students' prior scores. We can approximate these actual targets with a single coefficient, α, which multiplies a single prior score in the ABCs target, capturing the required test score growth rate under the ABCs.[77]

The growth targets are then aggregated across all grade-subject pairs within the school to form a school-level growth score for every school in the state. Thus a school passes the ABCs when the sum of the differences between average and target scores across all grades is greater than zero. We write the formal condition as

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

where G_s stands for the highest grade served at a given school.[78]

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

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

[77] In practice, the ABCs program sets grade-specific targets using both prior mathematics and reading scores, and individual multiplicative coefficients for each. While this is a simplification, the main feature to preserve is the linearity.

[78] 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.

which consists of a sum of grade-specific averages (represented by the upper bars).[79] The key term is the difference, $(\gamma_1^e - \alpha)$, namely how the persistence rate of effort compares with the target growth rate legislated under the ABCs. This captures the extent to which the 2002-03 NCLB response changes 2003-04 ABCs incentives. 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 becomes 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, the change in ABCs incentives should lead schools to respond by adjusting contemporaneous effort.

Such responses need accounting for in order to separately identify the direct persistent effect of once-lagged effort on student test scores. To do so, we recognize 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,[80] captured by $(\gamma_1^e - \alpha) \sum_{g=3}^{G} \bar{e}_{g-1,s,02-03}$ after moving the summation through (E.4).

On that basis, when estimating the persistence of NCLB effort in 2002-03, we control for average school-level 2002-03 NCLB effort to account for distortions to ABCs incentives. We state this formally as an assumption.

Assumption E.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, $\bar{e}^N_{s,02-03} = \sum_{g=3}^{G} \bar{e}_{g-1,s,02-03}$, 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, it is worth pointing out that our strategy does not control for the responses directly; instead, we hold the effort response fixed by accounting (based on the reasoning above) for the primary variable that determines the

[79] To see how equation (E.3) implies equation (E.4), consider the academic year 2003-04 and note that, given our framework, the contemporaneous test score may be written as the sum of the counterfactual predicted score in that year (capturing the effect of all non-effort inputs), contemporaneous effort, the persistent effect of prior-year effort, and measurement error. Likewise, the lagged (2002-03) test score may be written as the sum of the predicted score in that year (capturing the effect of all non-effort inputs), contemporaneous effort, and measurement error. Making the relevant substitutions for contemporaneous and prior-year test scores in equation (E.3) yields

$$\sum_{g=3}^{G_s} \sum_{i \in g} \frac{y^C_{i,j,g,s,03-04} + e_{j(i,03-04)} + \gamma_1^e e_{j(i,02-03)} + \eta_{i,j,g,s,03-04} - \alpha(\hat{y}_{i,j',g-1,s',02-03} + e_{j(i,02-03)} + \epsilon_{i,j'g-1,s,02-03})}{N_{g,s,03-04}} \geq 0.$$

Taking grade-level means then results in equation (E.4).

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

effort response, working through its effect on the degree to which a school alters its probability of passing the ABCs.

E.III.iii The Estimating Equation

Given the test score technology in (E.2) and definition of the counterfactual predicted score ($y^C_{i,j,g,s,03-04}$), the main estimating equation is written:

$$y_{i,j,g,s,03-04} - y^C_{i,j,g,s,03-04} = \gamma^e_1 e^N(\pi_{i,02-03}) + \underbrace{\theta e^N(\pi_{i,03-04}) + \rho \bar{e}^N_{s,02-03}}_{e_{i,j,g,s,03-04}} + \nu_{i,j,g,s,03-04}. \quad \text{(E.5)}$$

Deducting the counterfactual predicted score from students' realized scores on the LHS allows us to isolate the three effort responses that are relevant from an estimation perspective. The first term on the RHS – 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, represents effort in 2003-04, which depends on the key persistence parameter of interest (given that incentive strength in 2003-04 is partly a function of effort carrying over from 2002-03). The third term, following Assumption D.6, adds average school effort from 2002-03 multiplied by the parameter ρ in order to control for ABCs effort incentives in 2003-04.[81] (The underbrace in (E.5) highlights the way that the second and third terms on the RHS are subcomponents of contemporaneous effort, $e_{i,j,g,s,03-04}$, in our formulation.)

[81] 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^e_1 and ρ are not confounded.

Appendix F Policy Analysis: Estimating the Output-Incentive Expenditure Mapping

In this appendix, we set out our strategy for estimating the output-incentive expenditure mapping described in Section VI.C in detail.

Key to our approach is the dynamic link between incentives under NCLB and the ABCs. This dynamic link is apparent from our main estimates, which show that the NCLB effort response decreased the likelihood of ABCs target attainment the following year. This is because the persistence rate of effort (γ_1^e) is estimated to be lower than the required rate of growth under the ABCs, given by α.[82] 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, similar to the prediction in Macartney (2016).

The strategy for estimating the output-incentive expenditure mapping consists of the four steps given in the main text:

Step 1 - Calculating Schools' Expected Financial Losses Under the ABCs

First, we calculate the degree to which school responses to NCLB lowered – as the estimates indicate – the probability of passing the ABCs, relative to a counterfactual scenario in which NCLB was not introduced. To do so, we calculate the probability of passing the ABCs in 2003-04 under two different scenarios: one in which NCLB never occurred (the benchmark) and a second scenario in which NCLB was introduced and teachers responded to the prevailing incentives in 2002-03. The difference in these passing probabilities multiplied by the ABCs per-teacher bonus payment gives the expected per teacher loss (in dollars) from responding to NCLB's introduction.

Our calculations of these probabilities draw heavily on the structure and estimates from above. In our discussion of the institutional details of the ABCs in Section E.III.ii, we established that a school passes the ABCs in 2003-04 when the sum of the differences between average and

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

target scores (which are themselves a function of prior-year test scores) across all grades is greater than zero. This condition is given by equation (E.3) above, which we rewrite here for convenience:[83]

$$\sum_{g=3}^{G_s} \sum_{i \in g} \frac{y_{i,j,g,s,03-04} - \alpha y_{i,j',g-1,s',02-03}}{N_{g,s,03-04}} \geq 0.$$

Relying on the structure of our framework and the definitions in Table E.1 , we can write both the contemporaneous and prior score in this equation in terms of their respective component parts, which will facilitate our probability calculations.

Taking these in turn, the contemporaneous test score in 2003-04 is the sum of the counterfactual predicted score in that year ($y^C_{i,j,g,s,03-04}$, capturing the effects of all relevant non-NCLB effort inputs), contemporaneous effort ($e_{j(i,03-04)}$), the persistent effect of prior-year effort ($\gamma^e_1 e_{j(i,02-03)}$), and measurement error ($\eta_{i,j,g,s,03-04}$). Likewise, the prior-year test score from 2002-03 is the sum of the predicted score in that year ($\hat{y}_{i,j',g-1,s',02-03}$, reflecting the effects of all non-effort inputs), contemporaneous effort ($e_{j(i,02-03)}$), and measurement error ($\epsilon_{i,j'g-1,s,02-03}$). Making the relevant substitutions into equation (E.3), taking grade-specific means and rearranging yields the following passing condition:

$$\sum_{g=3}^{G_s} \left((\bar{y}^C_{g,s,03-04} - \alpha \bar{\hat{y}}_{g-1,s,02-03}) + \bar{e}_{g,s,03-04} + (\gamma^e_1 - \alpha)\bar{e}_{g-1,s,02-03} + \bar{\eta}_{g,s,03-04} - \alpha \bar{\epsilon}_{g-1,s,02-03} \right) \geq 0. \tag{F.1}$$

where the grade-specific means are indicated by the upper bars.

We use equation (F.1) to calculate the probability of a given school passing the ABCs in 2003-04 under each of the two scenarios we need to consider. Take first the benchmark scenario in which NCLB was not enacted in the prior year. We calculate the probability of a school passing the ABCs in this case by first using equation (F.1) but setting NCLB effort in both 2002-03 and 2003-04 equal to zero ($\bar{e}_{g,s,03-04} = \bar{e}_{g-1,s,02-03} = 0$) reflecting the fact that there were no NCLB incentives to respond to in this hypothetical scenario; in terms of our framework,

[83] Although we simplify the notation in this appendix for expositional ease, our calculations of ABCs growth scores follow the detailed rules for calculating school-level ABCs scores precisely. In particular, rather than take simple averages, we sum the weighted and standardized grade-and-subject-specific average differences between realized and target growth scores within each school.

a school passes the ABCs in 2003-04 under this 'no NCLB' scenario when

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

Both the average counterfactual predicted score ($\bar{y}^C_{g,s,03-04}$) and the average predicted score ($\bar{\hat{y}}_{g-1,s,02-03}$) are known quantities, calculated according the procedures outlined in Appendix E above. To calculate the probability that the condition given by equation (F.2) is satisfied for a particular school, we must make an assumption about the distribution of the stochastic component, $\sum_{g=3}^{G_s}(\bar{\eta}_{g,s,03-04}-\alpha\bar{\epsilon}_{g-1,s,02-03})$, reflecting the influence of noise at the school level.

Assumption F.1: the average test score noise in each school is distributed according to the cumulative density function $F(\cdot)$, represented using a normal distribution with mean zero and standard deviation σ.

We assess the sensitivity of our analysis to a variety of alternatives for σ.[84] We then calculate the probability that school s passes the ABCs in the 'no-NCLB' scenario as

$$1-F\left(-\sum_{g=3}^{G_s}\big(\bar{y}^C_{g,s,03-04}-\alpha\bar{\hat{y}}_{g-1,s,02-03}\big)\right). \tag{F.3}$$

Under the second scenario (allowing for effort in the first year NCLB was introduced), we wish to isolate the reduction in ABCs passing probabilities caused by the *initial* response to NCLB in order to determine the extent to which schools reduced their chances of passing the ABCs relative to the scenario without NCLB. To that end, we calculate ABCs school-level growth scores in 2003-04 while only allowing schools to respond with additional effort in 2002-03. We eliminate subsequent 2003-04 NCLB effort responses in the calculation by setting $\bar{e}_{g,03-04}=0$, in which case the passing condition from equation (F.1) becomes

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

To calculate the probability that this condition is satisfied for any given school, we again

[84]Specifically, we let the SD of school-level randomness vary from 0.1 to 1 developmental scale points in increments of 0.1 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 (between 0.01 and 0.1 for the standard deviation) does not appreciably alter any of our conclusions.

use the fact that both the average counterfactual predicted score ($\bar{y}^C_{g,s,03-04}$) and the average predicted score ($\bar{\hat{y}}_{g-1,s,02-03}$) are known quantities and substitute them into equation (F.4) directly. We calculate average effort in 2002-03 (given by $\bar{e}_{g-1,s,02-03}$) by first calculating an effort level for each student in that year using the student-specific effort function described in Appendix E above and then aggregating these quantities up to the school-grade level. Further, we can easily calculate $(\gamma^e_1 - \alpha)$, as we estimate $\gamma^e_1 = 0.10$, and the coefficient on the prior mathematics score under the ABCs ($\alpha = 0.68$) is fixed by the institutional arrangements that govern the program. Maintaining the same distributional assumption as in the first scenario above, the remaining stochastic component, $\sum_{g=3}^{G_s}(\bar{\eta}_{g,s,03-04} - \alpha\bar{\epsilon}_{g-1,s,02-03})$, is distributed according to the cumulative density function $F(\cdot)$, given by a normal distribution with mean zero. We then calculate the probability that school s passes the ABCs in 2003-04, given its response to NCLB in 2002-03 as

$$1 - F\left(-\sum_{g=3}^{G_s}\left(\bar{y}^C_{g,s,03-04} - \alpha\bar{\hat{y}}_{g-1,s,02-03}\right) - (\gamma^e_1 - \alpha)\bar{e}_{g-1,s,02-03}\right). \tag{F.5}$$

Subtracting the probability of passing the ABCs in the 'no-NCLB' scenario from the probability given by equation (F.5) results in the degree to which each school s lowered its likelihood (in percentage-point terms) of passing the ABCs in 2003-04 because of its effort response to NCLB in 2002-03:

$$\begin{aligned}\Delta F_s = & -F\left(-\sum_{g=3}^{G_s}\left(\bar{y}^C_{g,s,03-04} - \alpha\bar{\hat{y}}_{g-1,s,02-03}\right) - (\gamma^e_1 - \alpha)\bar{e}_{g-1,s,02-03}\right) \\ & + F\left(-\sum_{g=3}^{G_s}\left(\bar{y}^C_{g,s,03-04} - \alpha\bar{\hat{y}}_{g-1,s,02-03}\right)\right).\end{aligned} \tag{F.6}$$

Panel (a) of Table 6 provides a summary of the school-level passing probabilities under the two scenarios. As mentioned above, we assess the sensitivity of our analysis to a variety of alternatives for the standard deviation of school-level average test score noise, given by σ. For a 0.1 scale-point standard deviation in noise ($\sum_{g=3}^{G_s}(\bar{\eta}_{g,s,03-04} - \alpha\bar{\epsilon}_{g-1,s,02-03})$), 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 therefore 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.

Preferred Estimate for the Standard Deviation of the School-Level Error Term

Prior to describing the remaining three steps involved in recovering the output-incentive expenditure mapping, we describe briefly how we determine our preferred estimate for the standard deviation of school-level average test score noise, which is 0.36 developmental scale points.[85]

To start, for each school s, write the average test score noise that appears in equations (F.2) and (F.4) as $\nu_s \equiv \sum_{g=3}^{G_s}(\bar{\eta}_{g,s,03-04} - \alpha\bar{\epsilon}_{g-1,s,02-03})$. This captures the difference in mean test score shocks across adjacent years, with the required growth coefficient under the ABCs, α, multiplying the prior-year average. While ν_s is in practice a weighted average (where the weights are determined by the ABCs' rules), we take it to be a simple average. Further, we assume that student-level test score noise follows a stationary process and is uncorrelated – both within-student over time and across students – within and across years. We also assume student test score shocks are drawn from the same distribution in each grade.

On that basis, we simplify and write $\bar{\nu}_s = \bar{\eta}_{s,03-04} - \alpha\bar{\epsilon}_{s,02-03}$ as a simple school-level difference of average student test score noise across adjacent years and calculate its variance as $\frac{\sigma_\eta^2}{N_s} + \alpha^2\frac{\sigma_\epsilon^2}{N_s}$, where N_s is the total number of students attending school s. Stationarity implies $\sigma_\eta^2 = \sigma_\epsilon^2 = \sigma^2$, so the variance is $\frac{\sigma^2}{N_s}(1+\alpha^2)$. Our estimate of σ^2 is 20.14 developmental scale points (from the maximum likelihood results in Table 4). The average school size (grades 3 to 5), denoted $N_s =$, is equal to 227 students and $\alpha = 0.68$ for grade 4 mathematics. Substituting these values into the variance expression and taking the square root yields the standard deviation of 0.36 developmental scale points, which is our preferred estimate.

Step 2 - Calculating the Change in Financial Incentives for a One-Unit Change in School Effort

Equation (F.6) establishes that the degree to which each school lowered its likelihood of passing the ABCs in 2003-04 depends on both the average effort it exerted ($e^N_{s,02-03} = \sum_{g=3}^{G_s}\bar{e}_{g-1,s,02-03}$) and the difference between the persistence rate of effort and the growth rate of the ABCs target ($\gamma_1^e - \alpha$) – see the interaction on the first row. Because we find (as noted) that effort persists ($\gamma_1^e = 0.1$) at a lower rate than the (known) growth rate of the ABCs target ($\alpha = 0.68$), a

[85] As a reference point, 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.

one-unit increase in school-level effort in 2002-03 lowers the likelihood of the school passing the ABCs in 2003-04, and correspondingly creates an expected financial loss for each school.

In this second step, we estimate the relationship between these financial incentives and one-unit change in prior NCLB school-level effort. This represents the impact of an additional unit of effort on schools' subsequent financial incentives, which we later combine with the relationship between an additional unit of effort and subsequent student test scores in order to back out the direct effect of NCLB-induced changes in financial incentives on test scores.

We estimate the magnitude by which additional school-level effort in 2002-03 affected ABCs financial incentives in 2003-04 by regressing the expected dollar value each school stood to lose in 2003-04 (given by equation (F.6)) on average school-level effort from the prior year:

$$750 \cdot \Delta F_s = \alpha + \beta \bar{e}^N_{s,02-03} + \nu_{s,03-04}. \tag{F.7}$$

The estimate $\hat{\beta}$, which equals the response $\frac{d(750 \cdot \Delta F_s)}{d\bar{e}^N_{s,02-03}}$, governs the magnitude by which a one-unit increase in school-level effort in 2002-03 lowers the expected dollar amount the average school stood to gain under the ABCs financial incentives in 2003-04.

Panel (b) of Table 6 reports the estimated coefficients from equation (F.7). The coefficient ranges from -245.31, when the standard deviation of school-level noise is assumed to be 0.1 to -64.09 when the standard deviation is assumed to be 1. A one-unit increase in average school-level effort in 2002-03 thus results in an expected financial loss between \$64 and \$245 per teacher.

Step 3 – NCLB's Effect on Subsequent Student Test Scores

Next, in order to link passing probabilities (and the expected financial losses incurred) to scores, we need the effect of a one-unit change in average school-level effort from 2002-03 on test scores in 2003-04. This is captured by the parameter estimate $\hat{\rho} = 0.29$, already recovered in Section V.

The parameter can be written explicitly as the effect of lagged effort on scores: $\hat{\rho} = \frac{dy}{d\bar{e}^N_{s,02-03}}$. It reflects the indirect relationship between financial incentives under the ABCs and the corresponding teacher effort responses in 2003-04.

Step 4 – Relationship between Test Scores and Changes in Financial Incentives:

We now scale the effect of lagged school-level NCLB effort on ABCs financial incentives in 2003-2004 ($\hat{\beta} = \frac{d(750 \cdot \Delta F_s)}{d\bar{e}^N_{s,02-03}}$) by the effect of lagged school-level NCLB effort on test scores ($\frac{dy}{d\bar{e}^N_{s,02-03}} = \hat{\rho} = 0.29$). The relevant division 'cancels' the effort terms, giving the direct effect of test score gains on financial incentives, $\frac{\hat{\beta}}{\hat{\rho}} = \frac{750 \cdot d(\Delta F_s)}{dy}$. This represents the financial cost (in terms of bonus payment incentives offered) associated with a one-unit (one developmental scale point) gain in student test scores.

Panel (d) of Table 6 reports the direct effect of financial incentives on test scores, indicating that a one-unit increase in test scores is associated with a per-teacher cost ranging between $220 and $845. Our preferred estimate implies a cost of $504 per teacher.