Education Production and Incentives*

Hugh Macartney† Robert McMillan‡ Uros Petronijevic§

December 31, 2015

Abstract

The substantial 'value-added' literature that seeks to measure the overall impact of teachers on student achievement does not distinguish between teacher effects that are invariant to prevailing incentives and those that are responsive to them. In contrast, we develop an empirical approach that, for the first time, allows us to separate out incentive-varying teacher *effort* from incentive-invariant teacher *ability*, and further, to explore whether the effects of effort and ability persist differentially. Our strategy exploits exogenous variation in the incentive strength of a well-known federal accountability scheme, along with rich administrative data covering all public school students in North Carolina. We separately identify teacher effort and teacher ability to determine their relative magnitudes contemporaneously, finding that a one standard deviation increase in teacher ability is equivalent to 21 percent of a standard deviation increase in student test scores, while an analogous change in teacher effort accounts for 8 percent of such an increase. We then use prior incentive strength to reject the hypothesis that the persistence of teacher ability and effort is similar. To supplement our regression-based evidence, we set out a complementary structural estimation procedure, showing that effort affects future scores less than ability. From a policy perspective, our results indicate that incentives matter when measuring teacher value-added. Our analysis also has implications for the cost effectiveness of sharpening incentives relative to altering the distribution of teacher ability across classrooms and schools.

Keywords: Education Production, Value-added, Incentives, Accountability, Teacher Effort, Teacher Ability, Persistence

*We would like to thank Pat Bayer and Aloysius Siow for helpful comments and suggestions. Mike Gilraine provided outstanding research assistance. Financial support from the University of Toronto is gratefully acknowledged. All remaining errors are our own.

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

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

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

I. Introduction

Assessing the importance of teachers in the production of student achievement has been a central and long-standing preoccupation in academic research. Given that observable teacher characteristics tend to do a poor job of predicting student outcomes, several influential studies (for instance, Rivkin, Hanushek and Kain 2005) have used fixed effects methods to show convincingly that 'teacher quality' matters. This finding has led to the development of a substantial related literature, estimating sophisticated value-added measures that seek to capture the overall performance impact of a given teacher (see influential papers by Chetty, Friedman and Rockoff 2014a,b). These types of measure have become the focus of policy interventions that include, controversially, firing low value-added teachers, in turn leading to considerable scrutiny of the way value-added measures are constructed.

Alongside the research estimating teacher value-added, a second active literature has studied the impact of accountability incentives on student and school performance. Accountability schemes have become increasingly prevalent in the United States over the past two decades, sharing the common goal of increasing measured teacher and school outcomes by strengthening performance incentives. They range from the federal No Child Left Behind Act of 2001 ('NCLB'), a proficiency system that sets achievement targets and penalties for target non-attainment, to various state-level schemes setting rewards or imposing penalties based on the level or growth of student test scores. Persuasive evidence that accountability schemes succeed in improving student achievement in a variety of settings now exists.[1]

This paper brings these two literatures together for the first time. We extend the value-added literature by drawing a distinction between teacher effects that are invariant to the incentive environment and those that are responsive to it; for convenience, we label the former 'teacher ability' and the latter, 'teacher effort.' Our approach draws upon incentive variation associated with reforms analyzed in the accountability literature, setting out a new method for distinguishing ability from unobserved effort. In doing so, we show how value-added measurement can be refined to account for the prevailing incentive environment, offering an expanded set of policy levers to improve teacher and school performance.

The starting point for our analysis is a simple model of education production that captures the cumulative nature of student learning. It is natural (though quite rare in the literature) to allow for an effort margin on the part of educators in the student learning

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

process. Accordingly, we treat teacher effort and teacher ability as separate inputs to that process. Given that inputs may have a less-than-fully persistent effect on future learning outcomes, we also allow teacher ability and effort to persist at potentially different rates. While inputs may interact non-linearly, we focus on an additive specification – one that is likely to provide a reasonable first-order approximation to the true education production function. Under accountability incentives, teacher effort serves as a choice variable, and we draw attention to cases in which incentives are likely to be particularly strong – where educators are teaching marginal students, relevant for identification in the empirics.

This model forms the basis of a two-part empirical approach. In the first part, we exploit plausibly exogenous incentive variation to distinguish teacher ability from teacher effort in each period. Such variation is obtained from the introduction of NCLB in North Carolina. As is well-appreciated (see Reback 2008, for instance), proficiency schemes like NCLB make students matter differentially at the margin according to their likely test performance relative to a fixed performance target. Accordingly, our strategy entails comparing the teachers of students who are more versus less marginal, both before and after the incentive reform was implemented.

To decompose teacher effects into incentive-invariant ability and incentive-varying effort, we draw upon an important result from the recent literature: Overall teacher-effect estimates are shown to be unbiased predictors of teachers' average impact on student test scores when controlling for the prior scores of students, since non-random sorting of students to teachers occurs primarily with respect to that achievement measure. This finding is based on both experimental and quasi-experimental evidence, which indicates that only a few years of performance data are required in order to reasonably predict a teacher's value-added over her career (see Kane *et al.* 2013, Chetty *et al.* 2014a, and Kane and Staiger 2014).

We show that teacher-year effects of the type traditionally estimated do in fact vary with incentives. Further, our approach allows us to separate out the relative short-run importance of teacher ability and effort: we find that a one standard deviation increase in teacher ability is equivalent to 21 percent of a standard deviation increase in student test scores, while an analogous change in teacher effort accounts for 8 percent of such an increase. Given the level of incentives under the NCLB system, the associated increase in effort is substantial.

In the second part of our empirical approach, we use the contemporaneous measures of teacher ability and effort to investigate the extent to which each input persists (potentially differentially) in explaining future test scores. This component of the analysis is novel in that

persuasive evidence of the long-run effects of conventional teacher quality measures found in the literature (see Chetty *et al.* 2014b) does not allow a role for incentives.

We begin by estimating an interacted specification to test whether the coefficient on the prior score in a regression with the current score as the dependent variable differs significantly on a triple-differences basis: we difference across more and less marginal students, those previously taught in high- and low-effort classes, and those taught before and after the accountability reform's introduction. The findings from this interacted specification demonstrate that teacher ability and effort do not persist at similar rates: the triple-differences estimate for both grade three and four are positive and significant, while a placebo exercise using only pre-reform observations yields results that are statistically indistinguishable from zero.

The positive estimates we find using variation surrounding the introduction of the reform are indicative of effort persisting either at a lower or higher rate than ability and prior baseline effort. To determine which, we appeal to a complementary structural procedure, which utilizes information from both reduced-form portions of our empirical approach. It reveals that incentive effects fade at a faster rate than the combined effect of teacher ability and prior baseline effort after one year. This is in line with teachers 'teaching-to-the-test,' a phenomenon that is often discussed but rarely identified empirically.

These findings are relevant for policy. The prior literature (notably Hanushek 2011 and Chetty *et al.* 2014b) has focused on altering the teacher ability distribution as a policy lever to raise student scores. Specifically, policy proposals have featured the controversial notion of replacing teachers whose value-added falls in the bottom five percent of the measured distribution. Building on our finding that incentives matter when measuring the effects of teachers, changing formal incentives constitutes an alternative way of raising student and school performance. A straightforward procedure allows us to compare the benefit-cost ratios of these two rival types of policy intervention. Here, our analysis implies that targeting incentives to teachers in underperforming schools is likely to be more effective than firing teachers at the bottom of the value-added distribution if the cost of such an intervention is less than 17 percent of the latter policy in the long run.

The remainder of the paper is organized as follows: The next section sets out a simple framework for analyzing the production of student achievement that provides the basis for our empirical approach. Section III describes institutional background to the setting we consider, along with the rich set of data used in our analysis. Section IV outlines the

empirical strategies we employ to decompose ability and effort in both the short and long run, Section V presents our findings, and we interpret them in Section VI. Section VII draws out policy implications from the analysis and VIII concludes.

II. Theory

Central to our analysis is the technology governing education production. Education is a cumulative process that, in its essence, involves innate student ability interacting with parental and school inputs over time to determine student outcomes.[2] A completely flexible production technology would allow inputs to have differential effects on student outcomes depending on the grade (or age) of the student, the amount of time that elapsed since the inputs were applied, and the entire history of past inputs. As the exact specification of the technology is unknown and many inputs are unobserved, researchers typically limit analyses to linear production technologies and to inputs that are observed in school administrative data sets. Furthermore, most studies examine the effects of contemporaneous inputs, accounting for the history of past inputs by controlling for prior test scores.

Following the motivation above, we focus on both the contemporaneous and dynamic effects of two unobserved inputs, namely teacher ability and effort. As a starting point, we abstract from other inputs and consider the following production technology, which makes explicit how each of the two inputs affects student learning, both contemporaneously and over time:

$$y_{i,s(t),g(t),t} = \beta_{t,0} h_{i,0} + \sum_{0 \leq t' \leq t} \left[\gamma^a_{t,t'} a_{s(t'),g(t'),t'} + \gamma^e_{t,t'} e_{s(t'),g(t'),t'} \right] + \epsilon_{i,s(t),g(t),t} \,, \tag{1}$$

where $h_{i,0}$ is the human capital of student i at time 0, $a_{s(t),g(t),t}$ is the average ability of teachers at school s in grade g at time t, $e_{s(t),g(t),t}$ is the average effort of those teachers, and $\epsilon_{i,s(t),g(t),t}$ is an error term. To simplify further, let $\gamma^a_{t,t} = \gamma^e_{t,t} = 1$. Then, focusing on a student in cohort 0 (defined as $g = 0$ when $t = 0$), the equation for kindergarten ($g = 0$) simplifies to

$$y_{i,s(0),g(0),0} = h_{i,0} + a_{s(0),g(0),0} + e_{s(0),g(0),0} + \epsilon_{i,s(0),g(0),0} \,,$$

using $\beta_{0,0} = 1$. Analogously, the technologies for grades one, two, and three for a student who neither skips nor is retained a grade can be written as

[2] Individual student effort and the characteristics and behavior of other students matter also.

$$y_{i,s(1),g(1),1} = \beta_{1,0} h_{i,0} + a_{s(1),g(1),1} + e_{s(1),g(1),1} + \gamma^a_{1,0} a_{s(0),g(0),0} + \gamma^e_{1,0} e_{s(0),g(0),0} + \epsilon_{i,s(1),g(1),1} \,,$$

$$y_{i,s(2),g(2),2} = \beta_{2,0} h_{i,0} + a_{s(2),g(2),2} + e_{s(2),g(2),2} + \sum_{0 \leq t' < 2} \left[\gamma^a_{2,t'} a_{s(t'),g(t'),t'} + \gamma^e_{2,t'} e_{s(t'),g(t'),t'} \right] + \epsilon_{i,s(2),g(2),2} \,,$$

$$y_{i,s(3),g(3),3} = \beta_{3,0} h_{i,0} + a_{s(3),g(3),3} + e_{s(3),g(3),3} + \sum_{0 \leq t' < 3} \left[\gamma^a_{3,t'} a_{s(t'),g(t'),t'} + \gamma^e_{3,t'} e_{s(t'),g(t'),t'} \right] + \epsilon_{i,s(3),g(3),3} \,.$$

Given that we do not observe scores prior to grade three, we rewrite the grade three equation in the following way:

$$y_{i,s(3),g(3),3} = \beta_{3,0} h_{i,0} + I_{i,3} + a_{s(3),g(3),3} + e_{s(3),g(3),3} + \epsilon_{i,s(3),g(3),3} \,.$$

where the prior investment term $I_{i,3} \equiv \sum_{0 \leq t' < 3} \left[\gamma^a_{3,t'} a_{s(t'),g(t'),t'} + \gamma^e_{3,t'} e_{s(t'),g(t'),t'} \right]$ subsumes all scholastic inputs to that point. We can similarly write the expression for grade four:

$$y_{i,s(4),g(4),4} = \beta_{4,0} h_{i,0} + \tilde{I}_{i,4-3} + a_{s(4),g(4),4} + e_{s(4),g(4),4} + \gamma^a_{4,3} a_{s(3),g(3),3} + \gamma^e_{4,3} e_{s(3),g(3),3} + \epsilon_{i,s(4),g(4),4}$$

and beyond, where $\tilde{I}_{i,4-3} \equiv \sum_{0 \leq t' < 3} \left[\gamma^a_{4,t'} a_{s(t'),g(t'),t'} + \gamma^e_{4,t'} e_{s(t'),g(t'),t'} \right]$. By inspection, we see that $\tilde{I}_{i,4-3} = I_{i,3} + \sum_{0 \leq t' < 3} \left[(\gamma^a_{4,t'} - \gamma^a_{3,t'}) a_{s(t'),g(t'),t'} + (\gamma^e_{4,t'} - \gamma^e_{3,t'}) e_{s(t'),g(t'),t'} \right]$.

Our goals are to separately identify both teacher ability and effort, $\{a_{s(t),g(t),t}, e_{s(t),g(t),t}\}_{t=3}^{T}$, and their effects on test scores, $\{\gamma^a_{t,t'}, \gamma^e_{t,t'}\}_{t' < t}$. Separate identification of teacher ability and effort requires moving beyond teacher fixed effects, which subsume both, and relying on plausibly exogenous variation to exert effort across teachers. It also requires observing a teacher in at least two different time periods in which she faces different effort incentives. We argue below that the North Carolina setting and data satisfy both requirements.

Central to our approach is the way the effort choice of educators is treated as being endogenous to the prevailing incentive scheme, with agents balancing greater rewards against the convex cost of effort. This implies an effort function that depends on the parameters of the incentive scheme and, under threshold targets, an incentive strength measure describing how marginal the educator is given exogenous circumstances, as is quite standard. To predict where effort might be highest, we propose a continuous incentive strength measure based on the incentives implied by a proficiency scheme such as NCLB. This is described more fully in Section IV.

Given that we are able to separately identify the components of the amorphous 'teacher

quality' concept found in the literature, our next goal is to investigate how the effects of each persist over time. If $\gamma^a_{t,t'} \neq \gamma^e_{t,t'}$ for some $t' < t$, then ability and effort persist at different rates. Such a differential would have implications for the way in which student outcomes could be improved through rival policy reforms that alternatively alter the teacher ability distribution or incentive environment.

III. Institutional Background and Data

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

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

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

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

Our sample runs from 1997-2005 and is substantial, covering over 5 million individual-year observations. In terms of performance measures, it includes end-of-grade (EOG) test score performance data for mathematics and reading for all third to eighth grade public school students in the state. We also observe a 'pre-test' in grade three, which is written at the beginning of the year and treated as the grade two baseline test for third graders.

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

The longitudinal nature of the data set allows us to construct growth score measures for both mathematics and reading, based on within-student gains. These gains are positive, on average, in both subjects across grades, though the largest gains occur for both subjects in the earlier grades. Student gain scores are, as noted, the focus of the ABCs program, which sets test score *growth* targets for schools, requiring that students demonstrate sufficient improvement as they progress through their educational careers.

The data set includes information about individual students' gender, race, disability status, limited English-proficiency classification, free lunch eligibility, and grade progression. In the aggregate, about 39 percent of students are minorities (non-white), 14 percent are learning-disabled, only 3 percent are limited English-proficient, and 42 percent are eligible for free or reduced-price lunch. Around 27 percent of students have college-educated parents, and very small fractions of students skip a grade. These demographic characteristics serve as control variables in our analysis below.

Given our interest in exploring the separate effects of teacher ability and effort, we need to match students in the EOG files to their teachers in an accurate way in any given year. We construct the sample used in our analysis by following previous studies that use the NCERDC data, restricting attention to students in third through fifth grade, where the teacher recorded as the test proctor tends to be the teacher who taught the students throughout the year. We also follow Clotfelter *et al.* (2006) and Rothstein (2010) by only counting a student-

Table 1: Student-Level Summary Statistics

	Mean	Std. Dev.	N. Obs.
Performance Measures			
Math Score			
Grade 3	144.67	10.67	905,913
Grade 4	153.66	9.78	891,973
Grade 5	159.84	9.38	888,469
Grade 6	166.43	11.12	892,090
Grade 7	171.61	10.87	884,290
Grade 8	174.76	11.63	860,629
Math Growth			
Grade 3	15.77	15.30	842,656
Grade 4	9.38	5.97	741,220
Grade 5	6.82	5.30	741,009
Grade 6	7.51	5.70	739,810
Grade 7	5.96	5.64	736,029
Grade 8	3.74	5.89	715,966
Reading Score			
Grade 3	147.03	9.33	901,236
Grade 4	150.65	9.18	887,154
Grade 5	155.79	8.11	883,689
Grade 6	156.79	8.85	889,455
Grade 7	160.30	8.19	882,294
Grade 8	162.79	7.89	859,108
Reading Growth			
Grade 3	8.19	6.72	838,272
Grade 4	3.79	5.58	736,026
Grade 5	5.61	5.22	735,711
Grade 6	1.62	5.01	735,523
Grade 7	3.78	4.96	733,458
Grade 8	2.77	4.66	714,092
Demographics			
College-Educated Parents	0.27	0.44	5,456,063
Male	0.51	0.50	5,505,797
Minority	0.39	0.49	5,502,538
Disabled	0.14	0.35	5,497,987
Limited English-Proficient	0.03	0.16	5,505,597
Repeating Grade	0.02	0.13	5,505,797
Free or Reduced-Price Lunch	0.42	0.49	3,947,188

Notes: Summary statistics are calculated for all third through eighth grade student-year observations from 1997 to 2005. The free or reduced price lunch eligibility variable is not available prior to 1999. Math scores are measured on different scales before and after 2001. We are able to convert second edition scale scores to their first edition counterparts for all tests except the grade three pre-test (the grade two test). Thus, all level and gain math score summary statistics are expressed on the first edition scale except grade three gains, which are calculated using first edition scores prior to 2001 and second edition scores for 2001 onwards.

teacher match as valid if the test proctor in the EOG files teaches a self-contained class for the relevant grade in the relevant year and if at least half of the tests administered by that teacher are for students in the correct grade. Special education and honors classes are excluded from the analysis, but we retain students who repeat or skip grades.

IV. Empirical Approach

In this section, we describe our two-part reduced-form approach in some detail. To set the stage for that, we begin by placing it in the context of the existing teacher value-added literature.

IV.A. Teacher Value-Added

The recent teacher value-added (VA) literature has focused on credibly identifying the average causal effect of assigning a group of students to a given teacher. It is important to note, however, that teacher VA is not an input in the education production function, but rather a label given to systematic variation in the test scores of students who are assigned to a given teacher (see Jackson *et al.* 2014). The prior literature has therefore not been explicitly concerned with the way control function strategies interact with the underlying production technology that governs student learning; indeed, its primary concern has been the credible identification of the treatment effect of assigning students to a given teacher, which hinges on being able to account for non-random sorting of students to teachers adequately.

Our goal is different: we are interested in estimating inputs in the production technology and their effects. Since students and teachers are not randomly matched in our setting, we also adopt a control function strategy to account for systematic variation in student test score determinants across teachers. Given our goal of estimating the production technology in equation (1), however, it is important to clarify how estimates from a control function approach are interpreted, given our model.

Consider the production technology in equation (1), with all subscript notation augmented to the student level:

$$y_{i,s(i,t),g(i,t),t} = \beta_{t,0} h_{i,0} + \sum_{0 \leq t' \leq t} \left[\gamma^a_{t,t'} a_{s(i,t'),g(i,t'),t'} + \gamma^e_{t,t'} e_{s(i.t'),g(i,t'),t'} \right] + \epsilon_{i,s(i,t),g(i,t),t} \,. \qquad (2)$$

To see the result of controlling for the prior score, subtract α multiplied by the prior score

from both sides:[3]

$$y_{i,j,t} - \alpha y_{i,j,t-1} = \underbrace{a_{j(i,t)} + e_{j(i,t)}}_{q_{j(i,t)}} + \underbrace{(\beta_{t,0} - \alpha\beta_{t-1,0})h_{i,0}}_{\mu_{it}} + \cdots$$
$$\underbrace{\sum_{0\leq t'\leq t-1} \left[(\gamma^a_{t,t'} - \alpha\gamma^a_{t-1,t'})a_{j(i,t')} + (\gamma^e_{t,t'} - \alpha\gamma^e_{t-1,t'})e_{j(i,t')}\right]}_{\tilde{q}_{it}} + \underbrace{(\epsilon_{i,j,t} - \alpha\epsilon_{i,j,t-1})}_{\eta_{it}}. \tag{3}$$

The resulting reduced-form estimating equation, with the relevant relabeling, is

$$y_{i,j,t} = \alpha y_{i,j,t-1} + q_{j(i,t)} + \mu_{it} + \tilde{q}_{it} + \eta_{it}. \tag{4}$$

If the effect of initial student ability and previous teacher ability and effort do not all decay at the same rate α, $\mu_{it} + \tilde{q}_{it}$ will form part of the residual in this estimating equation along with η_{it}.

In our empirical approach below, we assume a teacher's ability and effort have a common effect on all of her students ($q_{j(i,t)} = q_{jt}$ for any student i assigned to teacher j in year t), as is standard in the literature. We estimate a more general version of equation (4), using teacher-year fixed effects to capture the combined impact of teacher ability and effort, q_{jt}. We then decompose the teacher-year fixed effect estimates into an incentive-invariant ability component and an incentive-varying effort component by exploiting plausibly exogenous variation in incentives to exert effort across teachers. Having separately identified ability and effort, we also investigate whether the effects of these inputs persist at different rates. This general background in place, the following two subsections describe the details of our two related empirical strategies.

IV.B. Identifying Ability and Effort

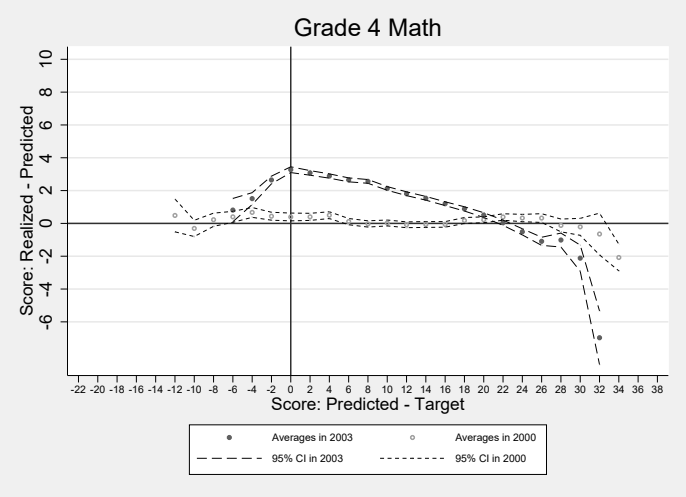
In this subsection, we set out our empirical strategy for decomposing a teacher's contribution to her students' test scores into ability and effort components.

In related work (Macartney *et al.* 2015), we show that the introduction of NCLB in North Carolina in 2003 had remarkable effects on students' high-stakes test scores, arguing that the observed test score improvements represent a teacher *effort* response. In this paper, we

[3]For convenience, we suppress the 'school-grade' subscript notation. In what follows, j denotes the teacher of student i at time t. The notation $j(i,t)$ makes that linkage explicit.

build on that earlier work, using the results to explicitly decompose teacher effects in 2003 into teacher ability and effort. We elaborate only on the details of Macartney *et al.* (2015) relevant for understanding our identification strategy here.

As noted in the introduction, it has already been well-established that proficiency-count systems like NCLB provide educators with strong incentives to direct resources to students who are on the margin of passing, potentially at the expense of those in the tails of the predicted test-score distribution.[4] Figure 1 shows that this was exactly the response that occurred in North Carolina when NCLB was introduced in 2003. The figure shows that gains over predicted test scores in 2003 were highest for students who were predicted to score close to the test-score proficiency threshold. Gains over predicted scores are decreasing as one moves farther away from the threshold, consistent with the predicted theoretical effort responses to proficiency-count programs. Importantly, students experience virtually no gain over their predicted scores at any point of the predicted score distribution in the pre-NCLB year, lending credence to the notion that the 2003 pattern reflects teachers' effort responses.



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

Figure 1: Inverted-U Response to NCLB

We use these results to separate teacher-year fixed effects in 2003 into more meaningful

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

components. Our approach involves a four-step procedure. By way of broad overview, we first compute teacher-year fixed effects for each teacher from 1997 to 2005. This step is largely in keeping with existing approaches taken in the literature. Second, we use pre-reform data to identify the sum of incentive-invariant ability and pre-existing baseline effort, given the teacher's fixed effect from step one and her experience profile. Third, we estimate NCLB-induced teacher effort, using estimated teacher fixed effects from 2003, estimates of the sum of teacher ability and baseline effort, and the fraction of students in a teacher's classroom who are 'marginal' with respect to the NCLB target.[5] Fourth, to include teachers who begin teaching in 2003 or later in our sample, we use effort estimates from step three for a given value of the proportion of marginal students in such a teacher's classroom to infer her combined level of ability and baseline effort. We now describe each step in greater detail.

1. Teacher-Year Fixed Effects: Recent VA studies typically control for high-order polynomials in previous math and reading scores along with students' demographic characteristics. We follow convention, regressing contemporaneous math scores on cubics in prior math and reading scores and several other student characteristics. While standardized (usually at the grade-year level) test scores are the predominant outcome measures in the literature, we opt to measure test scores on the developmental scale instead. Each approach has advantages and disadvantages. Standardizing test scores guards against changes in testing regimes over time but, importantly, de-meaning test scores within-year effectively removes much of the aggregate performance variation that is attributable to aggregate changes in performance incentives. As we are primarily interested in assessing how teacher effort affects student learning, we want to preserve much of the variation in test scores that is driven by aggregate changes in performance incentives over time. We thus keep test scores on the developmental scale throughout our analysis, relying on the careful psychometric design of these scales, which ensures that one can track improvements or declines in learning as students progress through school.

We define the control vector $x_{i,j,t}$ to include student race, gender, disability status, limited English-proficiency classification, parental education, and an indicator for grade repetition. These variables do not appear as test score determinants in the technology above. Their inclusion helps mitigate the bias stemming from non-random student-teaching matching, as they are likely correlated with innate student ability and previous teacher assignments. A teacher-year fixed effect is calculated for a teacher only if she has greater than seven but fewer

[5] We define a student as 'marginal' if she is predicted to score between four points below the proficiency threshold and four points above it.

than forty students in her class (in the relevant year) with valid test scores and demographic variables.[6]

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

$$y_{i,j,t} = f(y_{i,j,t-1}) + q_{jt} + x'_{i,j,t}\beta + \nu_{i,j,t}\,, \tag{5}$$

obtaining the teacher-year fixed-effects estimates as

$$\hat{q}_{jt} = \sum_{i=1}^{n(j,t)} \frac{\hat{y}_{i,j,t} - \hat{f}(y_{i,j,t-1}) - x'_{i,j,t}\beta}{n(j,t)} = a_j + \underline{e}_j + e_{jt}(a_j + \underline{e}_j, m_{jt}) + \bar{v}_{jt}\,, \tag{6}$$

where the second equality follows from equations (4) and (5) above. Note that $\underline{e}_j$ represents the teacher's baseline level of effort prior to the introduction of NCLB incentives, so that e_{jt} represents the additional effort induced by the NCLB reform. Thus, each fixed-effects estimate consists of incentive-invariant teacher ability, pre-existing effort, NCLB-based effort, and the component of $\tilde{q}_{it} + \eta_{it}$ that is orthogonal to controls.[7]

2. Incentive-Invariant Ability and Baseline Effort: Consistent with much of the existing literature, we assume incentive-invariant ability and baseline effort is fixed over time, conditional on teacher experience.[8] For each teacher j, we estimate her fixed ability by regressing the teacher-year fixed effects from the first step, $\hat{q}_{jt}$, on teacher and grade fixed effects, along with controls for teacher experience:[9]

$$\hat{q}_{jt} = a_j + \underline{e}_j + \lambda_g + h(exp_{jt}) + \zeta_{jt}\,. \tag{7}$$

[6]We mark a student as invalid for value-added analysis if he or she satisfies any of the following criteria: (1) multiple scores for current or lagged EOG math or readings tests; (2) EOG scores corresponding to two or more teachers in a given year; (3) EOG scores corresponding to two or more grades in a given year; (4) or EOG scores corresponding to two or more schools in a given year.

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

[8]For estimators that allow teacher ability to 'drift' over time, see Goldhaber and Hansen (2013), Chetty *et al.* (2014a), Rothstein (2014), and Bacher-Hicks *et al.* (2014). Assuming that baseline effort is fixed over time in our setting is reasonable, given that Macartney *et al.* (2015) finds that the pre-existing value-added incentive scheme is approximately uniform in its effects on teacher effort.

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

Prior to the introduction of NCLB, $e_{jt} = 0$. Therefore, the residual consists only of classroom shocks ϕ_{jt} which are not explained by teacher ability, base level effort, experience, or grade-specific determinants of performance ($\zeta_{jt} = \phi_{jt}$). Given such an error term and defining $\tilde{q}_{jgt} \equiv \hat{q}_{jt} - \hat{\lambda}_g - \hat{h}(exp_{jt})$, the estimate of a teacher's combined incentive-invariant ability and baseline effort is given by:

$$\widehat{a_j + \underline{e}_j} = \frac{1}{T} \sum_{t}^{T} \tilde{q}_{jgt} \,. \tag{8}$$

3. NCLB-Induced Effort Response: As previously mentioned, we define a student as 'marginal' if she is predicted to score within $+/-4$ developmental scale points of the proficiency cutoff. For each classroom, m_{jt} is defined as the fraction of students in that classroom who are marginal. Exploiting the descriptive relationship between teacher effort and a student's marginal status shown above, we parametrize the effort function in equation (6) to have the following functional form:

$$e(\widehat{a_j + \underline{e}_j}, m_{jt}) = [\pi_1(\widehat{a_j + \underline{e}_j}) + \pi_2]m_{jt} \,. \tag{9}$$

Once accountability pressure arrives in the form of m_{jt}, the teacher exerts additional effort, according to the amount of pressure she faces and her ability. Note that we cannot separate the base level of pre-reform effort from teacher ability, as our identification strategy depends on a shift in incentives. Thus, the effort we identify is that which occurs in response to the new incentive scheme only.

Under the NCLB reform, the residual from equation (7) becomes more complicated. It now reflects both the annual effort decision of teachers under NCLB e_{jt} and classroom shocks ϕ_{jt}, which are not explained by teacher ability, baseline effort, experience, or grade-specific determinants of performance. In particular, for $m_{jt} \neq 0$, we have that

$$\zeta_{jt} = \hat{q}_{jt} - \widehat{a_j + \underline{e}_j} - \hat{\lambda}_g - \hat{h}(exp_{jt}) = e_{jt} + \phi_{jt} = [\pi_1(\widehat{a_j + \underline{e}_j}) + \pi_2]m_{jt} + \phi_{jt} \,, \tag{10}$$

using the parametrization given by equation (9) and the estimated quantities from the first two steps.

The first year in which $m_{jt} \neq 0$ is 2003. Therefore, we estimate grade-specific regressions where we relate the residual given in equation (10) in that year to the fraction of students who are marginal in each classroom and the portion of ability that is not captured in equation

(7) when $m_{jt} = 0$ in the second step (recall $\zeta_{jt} = \phi_{jt}$ in that case):

$$\zeta_{j2003} = e_{j2003} + \phi_{j2003} = \pi_0 + \pi_1(\widehat{a_j + \underline{e}_j})m_{j2003} + \pi_2 m_{j2003} + \psi_{j2003} \tag{11}$$

In equation (11), systematic classroom shocks are contained in π_0, while unsystematic shocks are represented by ψ_{j2003}. Thus, we are able to identify the NCLB-induced effort response through variation in the fraction of a teacher's class that was marginal in 2003 and their ability. Our final estimates of each teacher's additional effort response is the predicted value from (11):

$$\hat{e}_{j2003} = \hat{\pi}_1(\widehat{a_j + \underline{e}_j})m_{j2003} + \hat{\pi}_2 m_{j2003}\,. \tag{12}$$

4. Post-Reform-Only Teachers: Given that the second step identifies the combination of teacher ability and baseline effort using pre-reform variation, a different technique is required to estimate that quantity for teachers who begin teaching in 2003 or later.[10] To obtain $\widehat{a_j + \underline{e}_j}$ for those teachers, we estimate a variant of equation (11) where $a_j + \underline{e}_j$ is unknown:

$$\zeta_{j2003} - \hat{\pi}_0 = e_{j2003} + \phi_{j2003} - \hat{\pi}_0 = \rho_1 m_{j2003} + \psi_{j2003}\,, \tag{13}$$

where $\hat{\pi}_0$ is estimated from equation (11). Assuming that the systematic component of classroom shocks for post-reform-only teachers is the same as for teachers with pre-reform experience, we have that $\hat{\rho}_1 = \hat{\pi}_1(\widehat{a_j + \underline{e}_j}) + \hat{\pi}_2$, which implies that $\widehat{a_j + \underline{e}_j}$ is given by

$$\widehat{a_j + \underline{e}_j} = \frac{\hat{\rho}_1 - \hat{\pi}_2}{\hat{\pi}_1}\,. \tag{14}$$

IV.C. Exploring Differential Persistence of Ability and Effort

Having separated teacher effort from teacher ability in the first part of our approach, we now exploit the differential responses across students to NCLB to test whether teacher ability and effort exhibit differential persistence in their effects on student future test scores.

A proficiency-count incentive system, such as NCLB, results in a much wider distribution of teacher effort across students than a value-added accountability system, such as North Carolina's ABCs. This follows from value-added programs using students' prior test scores

[10] Such teachers need to be included in the analysis to alleviate concerns about differential teacher turnover altering the post-NCLB teacher distribution and biasing the results.

to determine current performance targets. These schemes effectively set *student-specific* targets, which make all students marginal, providing educators with approximately uniform incentives across the student distribution.[11]

We use this feature of value-added schemes to argue that the effort students received prior to NCLB was essentially uniform – that is, students roughly received the same amount of effort from 1997 to 2002. In 2003, NCLB distorted teachers' decisions by providing incentives to direct more effort toward students who were predicted to score close to the proficiency target and to direct relatively less effort to students who were predicted to score farther away from it. The year 2003 thus provides a natural experiment in which the regime in North Carolina switches from one where all students receive the same amount of effort to one where marginal students receive a high amount of effort and non-marginal students receive a comparatively low amount of effort. To account for differential levels of effort within the +/- 4 band, we explicitly assume the following functional form for student-specific effort:

$$e^{\star}_{ijt} = \begin{cases} \frac{4-|\hat{y}_{ijt}-y^{T}_{ijt}|}{4} & \text{if } -4 \leq \hat{y}_{ijt} - y^{T}_{ijt} \leq 4 \\ 0 & \text{otherwise}\,, \end{cases} \tag{15}$$

where y^{T}_{ijt} is the proficiency threshold.

To identify γ^e, we exploit timing variation in the level of effort that teachers exert, using the 2003 NCLB effort shock for identification. Consider the model of student achievement from Section II. Assuming that γ^a and γ^e are each identical across time and grades, student achievement $\forall\ t \leq 2003$ is given by:

$$\begin{aligned} y_{i,j,t} = \alpha y_{i,j,t-1} + \mu_{it} + \sum_{0 \leq t' \leq t-1} \Big[[\gamma^a(1-\alpha)]^{t-1-t'} a_{j(i,t')} + [\gamma^e(1-\alpha)]^{t-1-t'} \underline{e} \Big] \\ + a_{j(i,t)} + \underline{e}_{j(i,t)} + \mathbf{1}_{t=2003} e_{j(i,t)} + \eta_{i,t}\,. \end{aligned} \tag{16}$$

For $t = 2004$, however, the equation differs slightly as effort deviates from base-level effort:

$$\begin{aligned} y_{i,j,t} = \alpha y_{i,j,t-1} + \mu_{it} + \sum_{0 \leq t' \leq t-1} \Big[[\gamma^a(1-\alpha)]^{t-1-t'} a_{j(i,t')} + [\gamma^e(1-\alpha)]^{t-1-t'} \underline{e} \Big] \\ + a_{j(i,t)} + \underline{e}_{j(i,t)} + e_{j(i,t)} + \gamma^e(1-\alpha) e_{j(i,t-1)} + \eta_{i,t}\,. \end{aligned} \tag{17}$$

[11] We explore these issues at some length in Macartney *et al.* (2015).

For $t = 2005$, we similarly have:

$$y_{i,j,t} = \alpha y_{i,j,t-1} + \mu_{it} + \sum_{0 \leq t' \leq t-1} \left[[\gamma^a (1-\alpha)]^{t-1-t'} a_{j(i,t')} + [\gamma^e (1-\alpha)]^{t-1-t'} \underline{e} \right]$$
$$+ a_{j(i,t)} + \underline{e}_{j(i,t)} + e_{j(i,t)} + \gamma^e (1-\alpha) e_{j(i,t-1)} + [\gamma^e (1-\alpha)]^2 e_{j(i,t-2)} + \eta_{i,t} \,. \quad (18)$$

Notice that the only difference between equations (16), (17) and (18) is the presence or absence of $\gamma^e(1-\alpha)$ terms multiplied by lagged effort levels $e_{j(i,t-t')}$. To create a valid counterfactual before and after NCLB (as $e_{i,j,t-1} = \underline{e}$ before NCLB), we generate estimates of NCLB-level effort for all years using the effort formula in equation (12):

$$\tilde{e}_{jt} = \hat{\pi}_1 (\widehat{a_j + \underline{e}_j}) m_{jt} + \hat{\pi}_2 m_{jt} \,. \quad (19)$$

We now have two additional layers of differences to exploit in addition to the pre- and post-NCLB comparison. First, differencing equation (17) from (16) reveals that the persistence of effort in the post-reform environment depends on the level of classroom effort $\tilde{e}_{i,j,t-1}$. Second, equation (15) also implies that effort may be directed disproportionately to particular students within the classroom ($e^{\star}_{i,j,t-1}$).[12] Thus, we estimate the following triple-differences specification to determine whether NCLB-induced effort decays differentially from the combination of teacher ability and baseline effort:

$$\begin{aligned} y_{i,j,t} = {} & \alpha_0 + \alpha_1 y_{i,j,t-1} + \alpha_2 \tilde{e}_{i,j,t-1} + \alpha_3 e^{\star}_{i,j,t-1} + \alpha_4 Post_t + \alpha_5 y_{i,j,t-1} \tilde{e}_{i,j,t-1} + \alpha_6 y_{i,j,t-1} Post_t \\ & + \alpha_7 \tilde{e}_{i,j,t-1} Post_t + \alpha_8 e^{\star}_{i,j,t-1} Post_t + \alpha_9 y_{i,j,t-1} e^{\star}_{i,j,t-1} + \alpha_{10} y_{i,j,t-1} \tilde{e}_{i,j,t-1} Post_t \\ & + \alpha_{11} y_{i,j,t-1} \tilde{e}_{i,j,t-1} e^{\star}_{i,j,t-1} + \alpha_{12} y_{i,j,t-1} e^{\star}_{i,j,t-1} Post_t + \alpha_{13} \tilde{e}_{i,j,t-1} e^{\star}_{i,j,t-1} Post_t \\ & + \beta_1 y_{i,j,t-1} \tilde{e}_{i,j,t-1} e^{\star}_{i,j,t-1} Post_t + \phi X_{i,j,t} + \lambda_{ct} + \eta_{i,j,t} \,, \end{aligned} \quad (20)$$

where $Post_t$ is a post-NCLB indicator, $X_{i,j,t}$ are student-level controls and λ_{ct} are classroom fixed effects. The triple-differences coefficient of interest is β_1. If $\beta_1 \neq 0$ then effort persists at a different rate than the combination of ability and baseline effort, since students who receive a high level of prior year classroom effort overall or a high amount of effort relative to their classmates would fare differently at time t than other students, conditional on their prior achievement level $y_{i,j,t-1}$.

Our specification controls for several student covariates and allows for the test scores of

[12] In practice, teacher effort is likely to be a superposition of common classroom effort and student-specific tutoring. We obtain identifying power from variation in both dimensions.

marginal students to persist differentially even in the absence of differential effort across students. For example, these students may have parents who encourage them to study more or less during the summer months, which would result in more or less knowledge transferring into future years. We also allow for the rate of persistence in prior scores to change for all students in the post-NCLB period. Conditional on prior test scores, our identifying assumption is that differences in the trajectory of achievement between marginal and non-marginal students across high- and low-effort classrooms and between pre- and post-reform observations only arise as a result of differential persistence.

V. Results

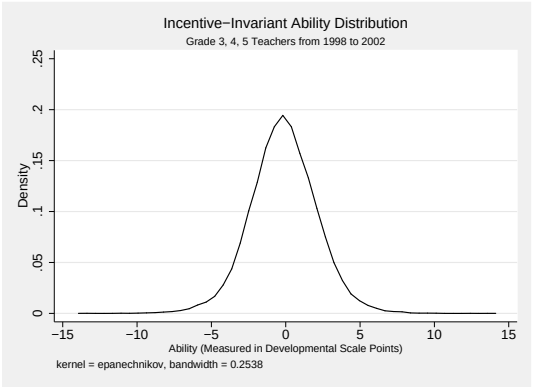
V.A. Identifying Ability and Effort

In this subsection, we present results relevant to the separate identification of teacher ability and effort. Figure 2 presents the incentive-invariant teacher ability distribution.[13] To be clear, incentive-invariant ability is defined as the teacher fixed effect from the regression in equation (7). The distributions are similar among third, fourth, and fifth grade teachers, with means of 0.14, -0.02 and -0.14 developmental scale points, and standard deviations of 2.21, 2.12 and 1.92 developmental scale points, respectively. Averaged across grades, mean teacher ability is 0.00 scale points and its standard deviation is 2.10 scale points, or equivalently 0.21 student-level standard deviations. The latter is within the upper end of the range found by most previous work for teacher quality overall.

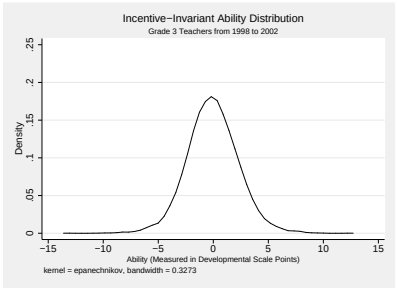
Figure 3 shows the 2003 effort levels obtained from the regression in equation (11). The top panels, (a) through (c), show the grade-year-specific estimated relationships between ζ_{jt} and m_{jt}. For each grade, we plot the relationships that prevail in 2003 and two control years. In 2003, there is a clear increasing relationship between the part of the teacher-year effect unexplained by incentive-invariant ability and the proportion of marginal students in the classroom. As expected, there is virtually no relationship in the pre-NCLB years, as the estimated functions are much flatter, or even downward-sloping in some cases.[14]

The solid dark lines in each of these panels represent the predicted effort values given by equation (11). To interpret the magnitude of the relationships in 2003, the slope coefficients in third, fourth and fifth grades are 1.83, 5.48, and 2.85, respectively. We take the fitted

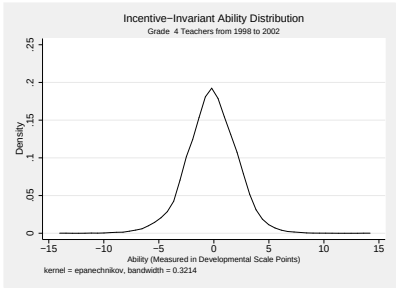
[13] This includes the baseline effort level discussed in the prior section, which is also invariant to NCLB incentives.
[14] The exception is in the year 2002 for fourth and fifth grades.



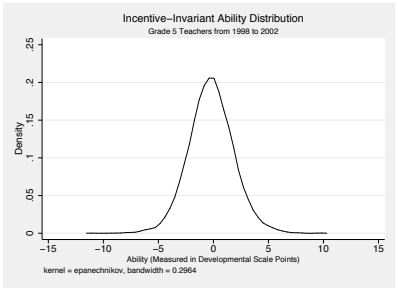
(a) All Grades



(b) Third Grade



(c) Fourth Grade



(d) Fifth Grade

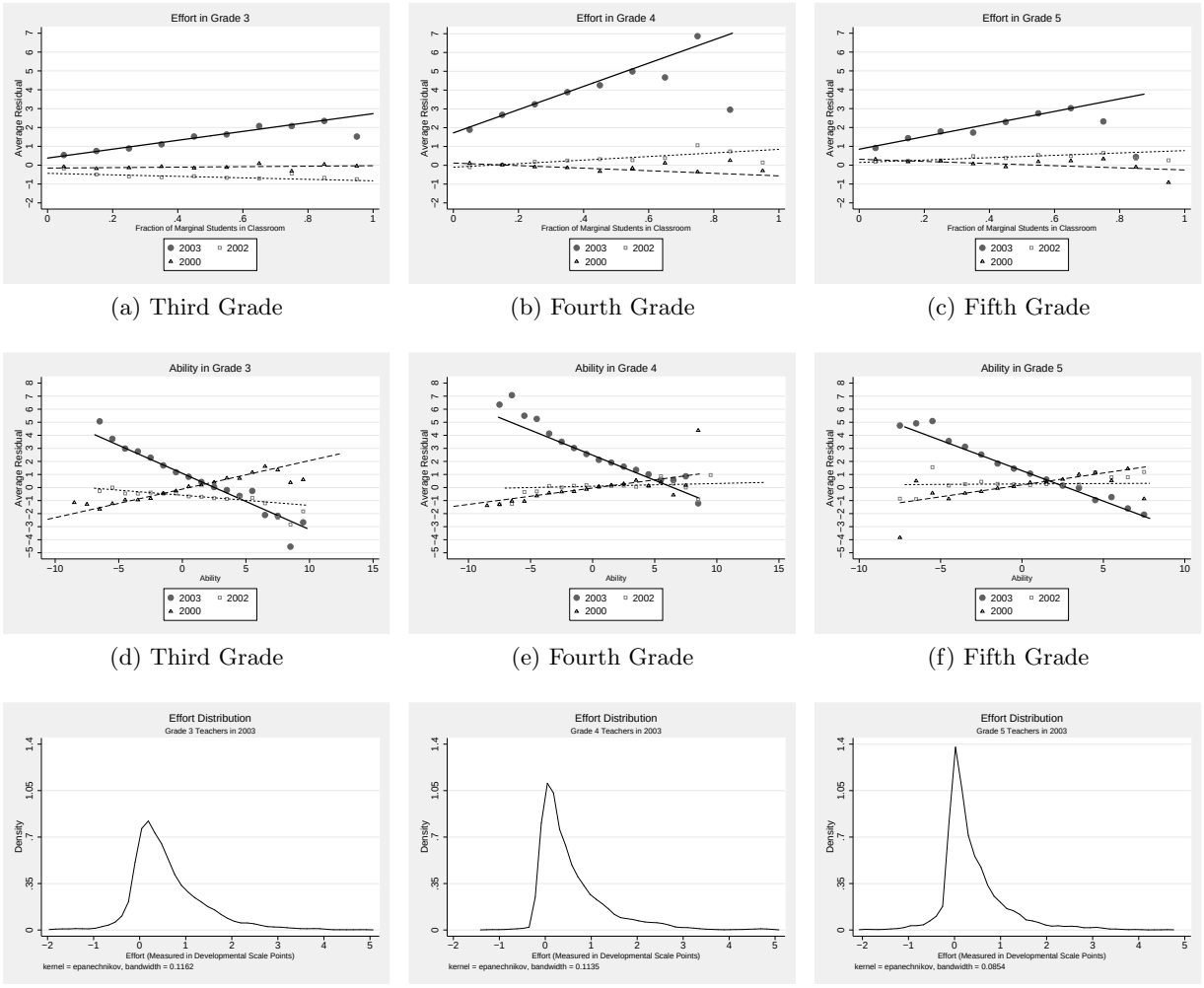
Notes: This figure shows the distributions of teachers' incentive-invariant abilities (which includes base level effort). To construct the figures, we estimate equation (7), having obtained the teacher fixed-effects from the regression in equation (5). Panel (a) shows the distribution of these fixed effects 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

values from this regression as the level of effort (conditional on ability) exerted by each teacher in 2003. Students in a classroom with a one standard deviation higher fraction of marginal students in 2003 (in their grade level) realized an average increase in their test scores of approximately 0.34 developmental scale points in third grade, 0.60 points in fourth grade, and 0.35 points in fifth grade. These improvements correspond to 0.04, 0.06, and 0.04 student-level standard deviations, respectively.

However, the reader may recall that we explicitly define effort as a function of both the proportion of marginal students in a teacher's classroom and the combined time-invariant ability and baseline effort component. Panels (d) through (f) reveal the grade-year-specific estimated relationships between ζ_{jt} and $a_j + \underline{e}_j$. For each grade, we plot the relationships that prevail in 2003 and two control years. In 2003, there is a clear decreasing relationship between the part of the teacher-year effect unexplained by incentive-invariant ability and base-level effort. In the control years, the estimated functions are upward sloping.[15] This

[15]Once again, year 2001 for grade 5 is an exception.



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

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

(g) Third Grade (h) Fourth Grade (i) Fifth Grade

Notes:This figure illustrates teachers' 2003 effort responses. In panels (a) to (c), we show the year-specific relationships between teacher-year residual test scores that are unexplained by incentive-invariant ability, teacher experience, and grade and time effects, and the fraction of students in a teacher's class who were marginal in each year. To construct these figures, we first obtain teacher-year residuals from equation (10). For each teacher-year, we then calculate the fraction of students in the class who were marginal (predicted to score with +/-4 points of the proficiency cutoff). The horizontal axis measures this fraction. We group teacher-year observations into 0.1-point width bins on the horizontal axis. Within each bin, we calculated the average teacher-year residual form equation (10). The circles in each panel represent these averages. The lines represent the associated linear effects, estimated on the underlying teacher-year data. Large dark circles and bold dark lines depict the relationship in 2003. The other circles and lines show pre-NCLB years. In panels (g) to (i), we present grade-specific densities of 2003 effort levels. To construct these figures, we first obtain 2003 effort for each teacher by taking the linear prediction (fitted value) from equation (12). We then plot the distributions of these effort levels separately by grade.

Figure 3: Effort Predictions and Distributions in 2003

suggests that it is the lower-ability teachers that respond to NCLB the most.

To interpret the magnitude of the relationships in 2003, the slope coefficients for third, fourth and fifth grades are -0.39, -0.33, and -0.43, respectively. We take the fitted values from this regression as the level of effort exerted by each teacher in 2003. Students in a classroom with a one standard deviation *lower* teacher ability in 2003 (in their grade level) realized an average increase in their test scores of approximately 0.80 developmental scale points in

third grade, 0.65 points in fourth grade, and 0.81 points in fifth grade. These improvements correspond to 0.08, 0.07, and 0.08 student-level standard deviations, respectively.

Panels (g) through (i) present the full distributions of effort in each grade in 2003. Mean effort for third, fourth and fifth grades is 0.61, 0.60, and 0.40 points, respectively (the average across grades is 0.54 points). Dispersion in effort is similar across grades (0.80, 0.80 and 0.74 points, respectively) is 0.79 scale points (or 0.08 student-level standard deviations). Taken together, the evidence suggests that there is meaningful variation in teachers' effort in 2003, driven by the fraction of students in their classes who are marginal.

V.A.1 Robustness Checks

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

A competing explanation is that students were *differentially* sorted to teachers in 2003, such that high (incentive-invariant) ability teachers received larger fractions of marginal students. Principals would be likely to sort students in this manner in response to NCLB if they believed the high-ability teachers had the best chance of achieving successful outcomes with these students. While we control for teacher fixed effects in the process of estimating effort responses in 2003, if high-ability teachers were better able to respond to the demands of NCLB, we might worry that this non-linear relationship between teacher ability and student ability is driving the results rather than additional effort being exerted by a *given* teacher.

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

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

If principals began differentially sorting students to teachers on the basis of ability in 2003, we would expect to find that $\beta_2 \neq 0$.

Table 2: Tests for Differential Sorting of Students to Teachers in 2003

	(1) Full Sample	(2) Third Grade	(3) Fourth Grade	(4) Fifth Grade
A: Incentive-Invariant Ability Estimated Using Full Sample				
Ability	-0.0057***	-0.0050***	-0.0057***	-0.0044***
	(0.0006)	(0.0013)	(0.0007)	(0.0012)
$1(t = 2003) \times$Ability	0.0007	-0.0024*	0.0029***	0.0014
	(0.0007)	(0.0013)	(0.0009)	(0.0010)
N	74,035	26,888	24,339	22,808
B: Teachers Before 2003				
Ability	-0.0057***	-0.0050***	-0.0057***	-0.0044***
	(0.0006)	(0.0009)	(0.0007)	(0.0012)
$1(t = 2003) \times$Ability	-0.0012	-0.0050***	0.0013	-0.0004
	(0.0008)	(0.0014)	(0.0010)	(0.0011)
N	68,646	24,759	22,607	21,280

Notes: This table presents the results of regressions based on equation (21). The dependent variable in each column is the fraction of students in a teacher's class who are marginal. In Panel A, teacher ability is estimated using the full sample from 1997 to 2005. In Panel B, we estimate teacher ability only for those who taught before 2003. Standard errors clustered at the school-level appear in parentheses. *** denotes significance at the 1% level;** denotes significance at the 5% level; * denotes significance at the 10% level.

Table 2 shows the results from estimating variants of equation (21). In Panel A, we use the full sample of teacher-year observations, by estimating incentive-invariant ability and effort by estimating for teachers that started in 2003 and beyond. Since there exists a positive relationship between teacher-year performance and the fraction of marginal students in a classroom in 2003, we might worry that the estimated incentive-invariant ability will capture some of this relationship, and that there will be a mechanical positive association between ability and the fraction of marginal students in 2003.

Columns (3) and (4) show that this appears to be the case for fourth and fifth grade teachers in 2003. The estimate on ability shows that high-ability teachers generally had lower fractions of marginal students in their classrooms: a 1 developmental scale point increase in teacher ability is associated with a 0.4 percentage point decline in the fraction of marginal students in the class.

In Panel B, we show that this is likely an artifact of the mechanical relationship described above. For each teacher that we observe in both the pre- and post-NCLB period, we then assign her the estimated teacher fixed effect from this regression as her ability measure. By construction, this measure does not pickup the association between the fraction of marginal students and teacher performance in 2003. When we use the pre-NCLB measure of ability,

there is no evidence of a positive change in the relationship between the fraction of marginal students and teacher ability in 2003 for fourth or fifth grade teachers. It does appear, however, that third grade students were differentially sorted in 2003 such that higher ability teachers received a *lower* fraction of marginal students than in previous years. This is sorting is in the opposite direction of that which is necessary for our results to be driven by teacher ability instead of effort.

To lend further credence to the strategy of using the proportion of marginal students in a classroom to identify effort responses, we relate the *improvement* in teacher-year effects from 2002 to 2003 to the fraction of marginal students teachers had in 2003. By making within-teacher comparisons in performance changes before and after 2003, we explore whether teachers with more marginal students in 2003 showed greater improvement.

We first construct the following difference for each teacher observed in 2002 and 2003:

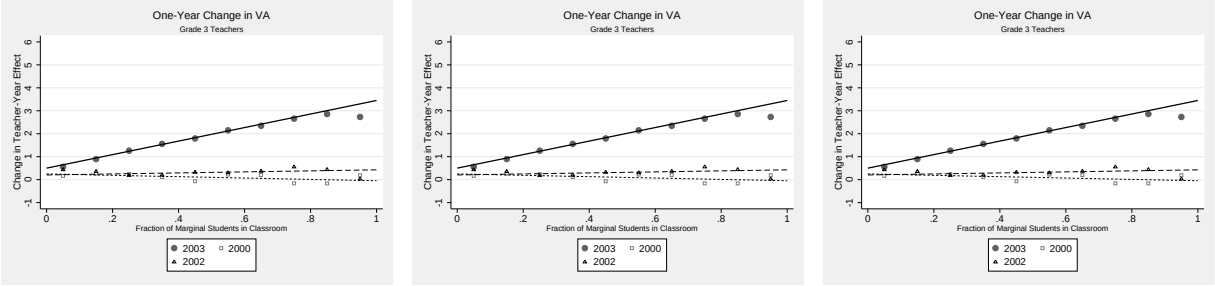
$$\hat{q}_{j2003} - \hat{q}_{j2002} = a_j + e_{j2003} + \bar{v}_{j2003} - (a_j + e_{j2002} + \bar{v}_{j2002}) = e_{j2003} - e_{j2002} + \bar{v}_{j2003} - \bar{v}_{j2002}, \quad (22)$$

and then estimate the relationship between this difference and the fraction of marginal students teachers had in 2003, m_{j2003}. If m_{j2003} is uncorrelated with changes in classroom-year specific shocks and $(\bar{v}_{j2003} - \bar{v}_{j2002})$ but is positively correlated with 2003 effort, we should see greater improvements in teacher quality for teachers with higher levels of m_{j2003}.

Figure 4 shows that the change in teacher quality is increasing in the fraction of marginal students teachers had in 2003. The control years in each grade reveal a flat relationship, which strengthens the claim that the 2003 patterns reflect teachers improving their performance as a result of NCLB incentives and the corresponding effort responses to these incentives.

V.B. Exploring Differential Persistence of Ability and Effort

In this subsection, we present the results of the test for differential persistence between ability and effort described in subsection IV.C. Table 3 below presents the results from estimating versions of equation (20). Panel A shows the results for fifth graders (who were in fourth grade when they were deemed either marginal or non-marginal) and Panel B presents the results for fourth graders (who were in third grade when they were deemed either marginal or non-marginal). Columns (2) and (4) show that students who were marginal in fourth grade generally have lower rates of persistence in fifth grade than students who were not



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

Notes: This figure depicts the relationship between the change in teachers' annual performance from 2002 to 2003 and the fraction of students in their classes who were marginal in 2003. To construct the figures, we first construct the change from $t-1$ to t between each teacher's teacher-year fixed effects from those years, as shown in equation (22). For each teacher-year, we then calculate the fraction of students in the class who were marginal (predicted to score with +/- 4 points of the proficiency cutoff). The horizontal axis measures this fraction. We group teacher-year observations into 0.1-point width bins on the horizontal axis. Within each bin, we calculated the average change in teacher-year fixed effects. The circles in each panel represent these averages. The lines represent the associated linear fits, estimated on the underlying teacher-year data. Large dark circles and bold dark lines depict the relationship in 2003. The other circles and lines show pre-NCLB years.

Figure 4: Changes in Teacher Quality by Fraction of Marginal Students within a Classroom

marginal, as the estimates in row (4) of Panel A are negative and significant for both the post (2004 and 2005) and pre (2001 and 2002) sample periods (as shown in Panel B, the finding also holds for fourth grade students in the pre-period). In column (5) of Panel A, the last row shows the estimate of β_1. It is highly significant and positive, suggesting that there is a difference between the persistence of NCLB-induced teacher effort and the combination of teacher ability and baseline effort for students transitioning to fifth grade. A similar result holds in Panel B for those transitioning to fourth grade.

Columns (6) and (7) conduct placebo tests. In column (6), we restrict the sample to 1999 and 2000, when there was no student accountability and define 2000 as the NCLB year of introduction; in column (7), we restrict the sample to 2001 and 2002, when there was student accountability, and define 2002 as the NCLB year of introduction. With the exception of the grade 3 to 4 transition in 2000, both columns reveal that there were no differential changes for marginal students across the two years, suggesting the estimates in column (5) reflect differential teacher, rather than student, effort.

VI. Structural Analysis

VI.A. Approach

In this section, we interpret the reduced-form estimates just presented, motivating the structural approach. That allows us to recover the persistence of both ability (including base

Table 3: Tests for Differential Persistence of 2003 Effort in 2004

	(1) 2004 & 2005	(2) 2004 & 2005	(3) 2001 & 2002	(4) 2001 & 2002	(5) Triple Differences	(6) 1999-2000 (Post=2000)	(7) 2001, 2002 (Post=2002)
				A: Grade 4 to 5 Transition			
y_{t-1}	0.782***	0.820***	0.707***	0.718***	0.720***	0.736***	0.689***
	(0.003)	(0.004)	(0.002)	(0.003)	(0.003)	(0.006)	(0.004)
$e^{\star}_{t-1}$		28.987***		18.871***	17.207***	23.985***	20.408***
		(1.813)		(1.433)	(0.914)	(2.338)	(1.769)
$\tilde{e}_{t-1}$		6.496***		0.595*	-0.391*	-0.922**	0.040
		(0.809)		(0.357)	(0.232)	(0.426)	(0.430)
$e^{\star}_{t-1} \times \tilde{e}_{t-1} \times y_{t-1}$		0.020**		-0.018***	-0.023***	-0.022***	-0.018***
		(0.009)		(0.004)	(0.003)	(0.006)	(0.006)
$Post \times e^{\star}_{t-1} \times \tilde{e}_{t-1} \times y_{t-1}$					0.044***	-0.015	-0.001
					(0.009)	(0.009)	(0.009)
N	102,009	87,420	112,025	89,869	312,144	85,231	89,869
R^2	0.765	0.771	0.784	0.796	0.791	0.796	0.797
	(1) 2004 & 2005	(2) 2004 & 2005	(3) 2001 & 2002	(4) 2001 & 2002	(5) Triple Differences	(6) 1999-2000 (Post=2000)	(7) 2001,2002 (Post=2002)
				B: Grade 3 to 4 Transition			
y_{t-1}	0.713***	0.698***	0.647***	0.663***	0.678***	0.726***	0.664***
	(0.003)	(0.005)	(0.002)	(0.003)	(0.002)	(0.004)	(0.004)
$e^{\star}_{t-1}$		10.794***		12.273***	14.911***	15.621***	14.919***
		(1.242)		(1.056)	(0.647)	(1.550)	(1.386)
$\tilde{e}_{t-1}$		3.871***		-0.363	0.986***	0.453	0.460
		(0.882)		(0.459)	(0.360)	(0.654)	(0.636)
$e^{\star}_{t-1} \times \tilde{e}_{t-1} \times y_{t-1}$		0.023***		-0.019***	-0.008**	-0.036***	-0.021**
		(0.008)		(0.006)	(0.003)	(0.008)	(0.008)
$Post \times e^{\star}_{t-1} \times \tilde{e}_{t-1} \times y_{t-1}$					0.031***	0.031***	-0.003
					(0.008)	(0.011)	(0.010)
N	101,565	76,201	118,312	87,897	292,749	82,590	87,897
R^2	0.742	0.756	0.767	0.783	0.776	0.771	0.783

Notes: The full triple-differences specification (column (5)) uses observations from the years 1998-2005 (2004 and 2005 are post-reform years), while the other columns specify which years are included. The dependent variable in each column is the EOG math score. Each regression additionally controls for student ethnicity, limited English proficiency, parental education, disability status, free or reduced-price lunch eligibility, gender, and class fixed effects. Standard errors clustered at the school level appear in parentheses. *** denotes significance at the 1% level;** denotes significance at the 5% level; * denotes significance at the 10% level.

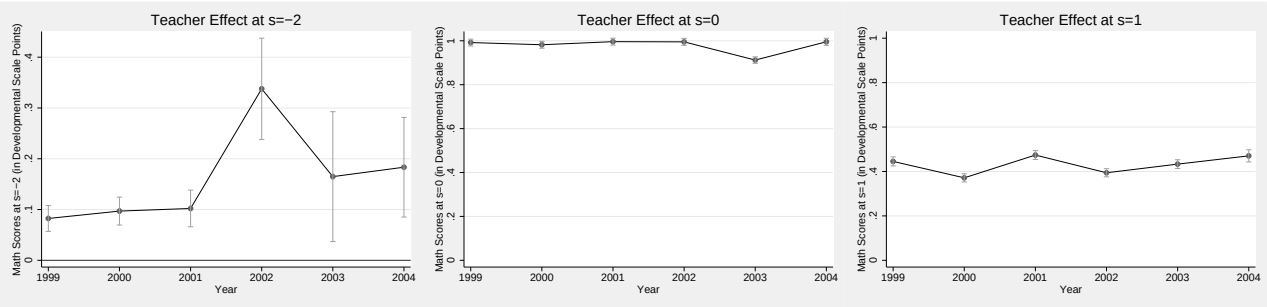
level effort) and incentive-based effort while allowing incentives to be correlated over time – an important feature of the incentive environment we consider.

Recall that Table 3 shows that the underlying decay rates of effort and ability differ. We now propose a structural method to recover estimates of both γ_a and γ_e. The basic steps for the structural estimation are: (1) find the teacher effect on student achievement in t and $t+1$; (2) using pre-2004 data when we know that there is no NCLB-incentive-based effort decay, use the difference in the teacher effect from t to $t+1$ to determine the persistence of ability; and (3) use the difference in the teacher effect from t to $t+1$ in 2004 and the persistence of ability from step (2) to identify the persistence of effort.

Following Chetty *et al.* (2014a), we run the following regression to recover the overall persistence of an increase in $\hat{q}_{j(i,g)}$:

$$y_{i,j,t+s} = \beta_{g,t+s}\hat{q}_{j(i,g)} + f(y_{i,j,t-1}) + h(exp_{jt}) + \phi X_{ijt} + \epsilon_{i,j,t} \tag{23}$$

where $s \in \{-2, 0, 1\}$. When $s = -2$, we identify the effect of teacher j on outcomes before she teaches student i.[16] Since the teacher has yet to actually teach the child, these estimates should always be zero. When $s = 0$, we recover the effect of the teacher on student achievement in that period. In contrast, $s = 1$ is the effect on student outcomes the following year; thus the difference between outcomes in $s = 0$ and $s = 1$ is the fade out of the teacher fixed effect.



(a) When $s = -2$ (b) When $s = 0$ (c) When $s = 1$

Notes: This figure shows the effect of teacher value-added on student outcomes by year, for outcomes two years prior (panel (a)), the current year (panel (b)) and the following year (panel (c)). To construct the figures, we estimate equation (23) and obtain the estimates for each year of the regression.

Figure 5: The Effect of Teachers in Years $s \in \{-2, 0, 1\}$

Figure 5 reports the results of equation (23) by year. From panel (a), notice that the

[16] When $s = -1$, the effect of teacher j is zero by construction as we are controlling for test scores in that year.

teacher-year effect does have some predictive power with respect to the outcomes of students two years before they are taught by that teacher. This is likely due to the mechanical correlation discussed in Chetty *et al.* (2014a).[17] For the effect of the teacher-year-effect on test scores in the current year, we see that the effect is stable around 1 developmental scale point until 2003, when it drops to 0.91 developmental scale points. Since this drop is not observed in other years, this leads to *a higher* level of persistence in 2003 relative to prior years. This may seem strange at first glance, as teacher effort increased in 2003 due to the introduction of NCLB. Consequently, one might expect persistence to be lower in 2003. However, due to the change in effort in 2003, a one-unit increase in $\hat{q}_{j(i,g)}$ may no longer correspond to the same increase in ability in 2002. Similarly, a one-unit increase in $\hat{q}_{j(i,g)}$ now corresponds to a change in effort post-2002. We estimate the following regressions to understand this relationship:

$$\hat{q}_{j(i,g),t} = \alpha_0 + \beta_{aq}\widehat{(a+\underline{e})}_{j(i,g),t} + \beta_{eq}\hat{e}_{j(i,g),t} + \epsilon_{i,j,t}\,. \tag{24}$$

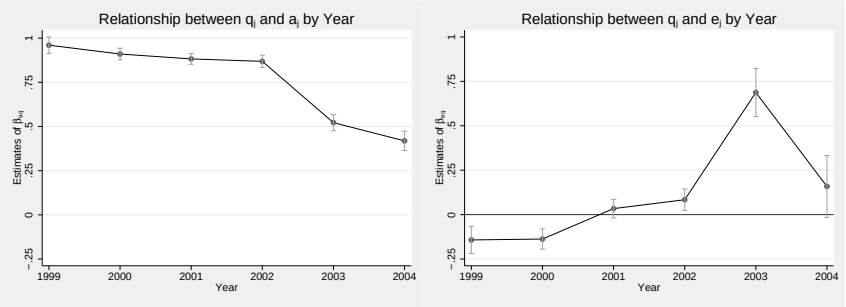
Estimates of β_{aq} and β_{eq} reveal how a one-unit increase in incentive-invariant ability (along with baseline effort) $\widehat{(a+\underline{e})}_{j(i,g),t}$ or NCLB-induced effort $\hat{e}_{j(i,g),t}$ maps into the teacher-year fixed effect $\hat{q}_{j(i,g)}$. In particular, $\hat{q}_{j(i,g)}$ increases by one unit if either $\widehat{(a+\underline{e})}_{j(i,g),t}$ increases by $\frac{1}{\hat{\beta}_{aq}}$ or $\hat{e}_{j(i,g),t}$ increases by $\frac{1}{\hat{\beta}_{eq}}$. The estimates are shown in panels (a) and (b) of Figure 6. Upon the introduction of NCLB in 2003, a one-unit increase in $\hat{q}_{j(i,g)}$ is associated with an increase of $\frac{1}{0.606} = 1.65$ units in incentive-invariant ability and $\frac{1}{0.708} = 1.41$ units in NCLB-induced effort.

We now model student achievement explicitly as a function of ability and effort. Specifically, we assume that ability and effort in year $s = 0$ generate achievement proportionally to the levels of $\hat{a}_{j(i,g)}$ and $\hat{e}_{j(i,g)}$ observed in that year. Therefore:

$$\hat{\beta}_{g,t} = \frac{\kappa_a}{\hat{\beta}_{aq}} + \frac{\kappa_e}{\hat{\beta}_{eq}} \tag{25}$$

where κ_a and κ_e relate how a one-unit increase in ability and effort increase achievement, respectively. Prior to 2003, we set $\hat{\beta}_{eq} = 0$ to reflect the fact that no NCLB-related effort should be exerted. To account for the systematic matching of low-ability teachers to classes with high levels of marginal students, we set $\hat{\beta}_{eq}$ from 2003 onward to the difference between

[17] A likely reason for such a mechanical correlation is that we are not leaving student i out of the estimation of $\hat{q}_{j(i,g)}$ at this time.



(a) Ability Component (b) Effort Component

Notes: This figure reveals how a one-unit increase in the teacher-year fixed effect is associated with an increase in both ability and effort.

Figure 6: Relationship Between Teacher-Year Fixed Effect and Its Components

$\hat{\beta}_{eq}$ in the pre- and post-NCLB periods (thus assuming that the matching function does not change in response to NCLB, which is justified by the evidence presented in Table 2).

The effects in $s = 0$ now persist at the rate of γ, where we allow for differential persistence for effort and ability, in the following year ($s = 1$). Therefore, achievement at that point in time can be written as:

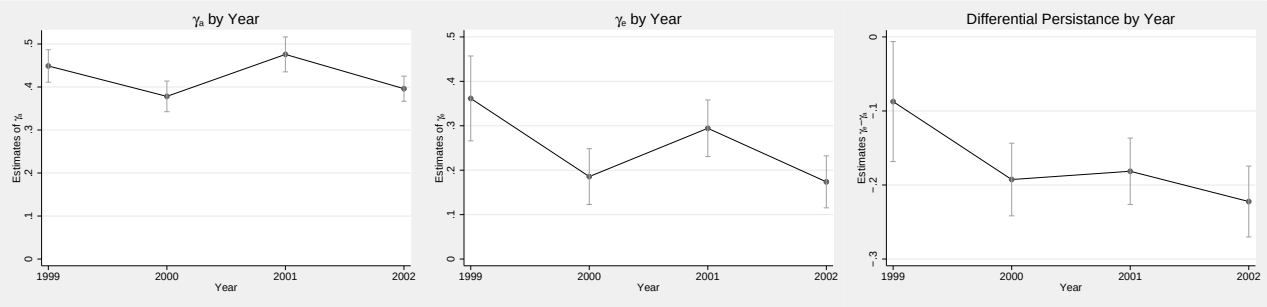
$$\hat{\beta}_{g,t+1} = \frac{\kappa_a \gamma_a}{\hat{\beta_{aq}}} + \frac{\kappa_e \gamma_e}{\hat{\beta_{eq}}} \tag{26}$$

Assuming constant γ_a and γ_e parameters, we can solve for each of them according to the following procedure: use pre-NCLB data (when $e = 0$) to find κ_a from equation (25) and γ_a from equation (26); then, given γ_a and κ_a, solve for γ_e using post-NCLB data.

VI.B. Persistence Estimates

Using pre-NCLB data, we estimate γ_a for each year prior to the introduction of NCLB. Figure 7(a) reports these estimates. As the figure reveals, $\hat{\gamma}_a$ is relatively constant over time, though it does rise somewhat in 2001. Given that estimating γ_e first requires an estimate of γ_a from the pre-NCLB data, we experiment with different pre-NCLB years to compute $\hat{\gamma}_e$ (reported in Figure 7(b)).

Figure 7(c) then shows the difference between $\hat{\gamma}_a$ and $\hat{\gamma}_e$. Table 4 reports the numbers underlying Figure 7. As expected from the reduced form estimates, we see that $\gamma_a > \gamma_e$, both for the year preceding the NCLB reform (2002) and when averaging across the four prior years (1999-2002). We take this as evidence of teachers 'teaching-to-the-test,' a phenomenon



(a) γ_a (b) γ_e (c) Differential persistence ($\gamma_e - \gamma_a$)

Notes: Panel A shows the estimates of γ_a by year, while panel B shows the estimates of γ_e using different base years. Similarly, panel C shows the difference in γ_a and γ_e using the different base years. Standard errors are clustered at the school level and are found using the delta method.

Figure 7: Estimates of Differential persistence by Base Year (1999-2002)

Table 4: Persistence Parameters by Year

Year:	2002	1999-2002
γ_a	0.396***	0.431***
	(0.015)	(0.013)
γ_e	0.174***	0.312***
	(0.030)	(0.024)
$\gamma_e - \gamma_a$	-0.222***	-0.119***
	(0.024)	(0.019)

Notes: This table presents estimates of γ_a and γ_e using different pre-NCLB years as the baseline for estimation. Standard errors are clustered at the school level and are found using the delta method. *** denotes significance at the 1% level.

that has been very difficult to credibly identify in prior work. Interestingly, given that $\gamma_e > 0$, we are able to show that the teacher effort response to NCLB is not fully transitory. Instead, it persists at between 44 and 72 percent of the rate of persistence for ability. This is important as it shows that effort, just like ability, matters not only contemporaneously but into the future as well.

VII. Policy Analysis

Our analysis is relevant to recent education policy discussions. The prior literature, notably Chetty *et al.* (2014b) and Rothstein (2015), focuses on altering the teacher ability distribution as a policy lever to raise student scores. Specifically, policy proposals have centered on the radical notion of firing teachers whose value-added falls in the bottom of the measured

distribution (for example, the bottom five percent).[18]

In contrast to such policies, an alternative method for improving student performance is to alter the teacher effort distribution. While existing estimates in the literature are unlikely to be informative for policymakers who are considering incentive reforms, our empirical framework and findings are particularly well suited to evaluating the cost effectiveness of rival ability- and effort-oriented policies. Based on our reduced-form and structural estimates, we calculate that incentive reform is more cost-effective in the short run if it costs less than 38 percent of the ability-targeted reform and, assuming that the persistence rates of ability and effort converge after one year,[19] incentive reform is preferred in the long run if it costs less than 17 percent of its ability-based counterpart.

To understand whether incentive policies are likely to pass this test, it is important to discuss the costs associated with the rival policies as well as to place them on equal footing in terms of benefits. Although our sample differs somewhat, Chetty *et al.* (2014b) estimates that replacing the lowest rated teachers with draws of new teachers would result in an average two standard deviation increase in ability for that subset. However, based on estimates from Rothstein (2015), doing so would require a mean salary increase across all teachers of 1.4 percent, as compensation for increased employment risk.

Our prior work (Macartney *et al.* 2015) suggests two avenues for increasing effort. First, policymakers can increase the proportion of marginal students by altering the target of non-marginal students. Non-marginal students with predicted performance above the NCLB accountability target account for about 60 percent of our sample, which implies that raising the proportion of marginal students in each class by 60 percent (from an average of 19 percent) would be a costless reform. Such a change in targets would increase effort by 0.2 standard deviations of the test score, which is equivalent to a one standard deviation increase in ability. The second avenue for increasing effort is to intensify the severity of the sanction under NCLB (or, equivalently, raise the amount of the reward under a comparable pecuniary scheme). Given that the proportion of marginal students in the average class would be 79 percent after implementing tougher targets (and the effect of a 60 percent increase in the proportion of marginal results is a one standard deviation increase in ability), raising the level of sanction (or reward) by 75 percent would result in a further one standard deviation increase in ability.

[18]The literature has also focused on reducing the attrition of the highest rated teachers. However, estimates suggest that such a focus on the top is a less cost-effective ability-oriented reform than replacing the lowest rated.

[19]We plan to test this hypothesis by structurally estimating persistence rates further into the future.

In combination, the two avenues for increasing teacher effort result in a benefit that is comparable to replacing the bottom five percent of teachers (in terms of value-added performance) with average draws. However, the effort-based benefit occurs for *all* teachers, while the ability-oriented reform only has an effect on five percent of the teacher distribution, at the cost of increased wages for all teachers. Consequently, sharpening incentives appears to compare favorably with replacing the lowest rated teachers in both the short and long run.

VIII. Conclusion

In this paper, we have presented a two-part strategy that permits us, for the first time in the literature, to separate out teacher effort, which is responsive to education incentives, from teacher ability, which is not. Further, we measure the extent to which these two potentially important education inputs persist differentially.

Central to our analysis was a novel identification strategy taking advantage of a natural experiment associated with the introduction of a federal accountability program in a setting – the state of North Carolina – where accountability incentives already operated. Specifically, we drew on the proficiency-count design of NCLB to construct a measure of incentive strength for each teacher, showing a positive relationship between teacher performance and this measure in the year NCLB was introduced but not in prior years. We exploited these differential relationships over time to separate teacher quality into ability and the effort response associated with NCLB.

We found greater dispersion in teacher ability than in NCLB-induced teacher effort, but both have significant effects on student achievement: a one standard deviation increase in teacher ability causes a 20 percent of a standard deviation increase in student achievement, while a one standard deviation change in effort leads to between a 3 and 5 percent of standard deviation change in achievement.

To evaluate the persistence of teacher ability and effort, we took advantage of the difference between the effort arising from the state-level ABCs and federal NCLB incentive schemes, comparing the knowledge persistence of students in danger of failing to achieve proficiency status before and after the introduction of NCLB. This evidence helps shed light on the extent to which 'teaching to the test' occurs in practice – a term often mentioned in the literature, but with little statistical evidence to indicate that it is important. Here, we

found that the effort received by fourth graders under NCLB is more transitory than the effort received by fourth graders under the ABCs, while the opposite result holds for third grade. These findings point toward the need to better understand the optimal timing of effort investments and the type of effort teachers exert across grades in response to NCLB.

We argued that, if effort investments in subsequent years are correlated with investments in prior years, as is likely, so our persistence estimates will reflect this correlation in addition to the true persistence of effort. The dynamic link between effort investments thus requires us to explicitly model effort decisions as a function of incentives in order to isolate contemporaneous teacher effort from prior investments. Doing so allowed us to isolate yearly measures of teacher effort consistently and to evaluate its persistence at each point in time.

Based on our estimates, we were able to explore the policy implications of rival education reforms. Our analysis suggests that using formal incentives constitutes a viable alternative means of accomplishing the goal of raising student and school performance, and is preferable to ability-based reforms in many instances. Overall, our findings indicate that the effort margin is first-order: teacher effort is both a productive input and one that is responsive to incentive variation in a systematic way.

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. 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review*, 104(9): 2593-2632.

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

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

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

Hanushek, Eric A. 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, C. Kirabo, Jonah Rockoff, and Douglas Staiger. 2014 "Teacher Effects and Teacher Related Policies." *Annual Review of Economics*, 6(34): 801-825.

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

Kane, Thomas J. and Douglas O. Staiger. 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. De-

signing teacher evaluation systems: New guidance from the Measures of Effective Teaching project. San Francisco.

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

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

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

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

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

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

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

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

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

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

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