The Consequences of Combining Apparently Unrelated Accountability Regimes*

Hugh Macartney† Robert McMillan‡ Uros Petronijevic§

September 12, 2014

*We would like to thank Aloysius Siow for helpful discussions. Financial support from Canada's Social Sciences and Humanities Research Council (SSHRC) is gratefully acknowledged. All remaining errors are our own.

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

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

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

Abstract

Accountability reforms seek to improve school performance by sharpening incentives. Over the past decade, many states have implemented both state-level and federal accountability programs with distinct, and sometimes conflicting, incentives. We analyze an environment in which educators in a state were first subjected to a sophisticated value-added incentive scheme for several years before also being held accountable under a supposedly unrelated and decidedly less efficient proficiency-based scheme. Using rich longitudinal data covering all public school students in the state, we show that the introduction of the proficiency-count program led to marked improvements in school performance – a surprising result, given the high-powered monetary reward system already in place. We further document potentially conflicting incentives between the two regimes, as the abrupt improvements in performance levels just referred to made it difficult for schools to satisfy their growth targets in future years, thereby jeopardizing their chances of earning monetary rewards under the pre-existing value-added program. We show that some of these improvements were generated by directing additional resources toward students in the middle of the achievement distribution – a well-documented response to proficiency-count systems but an unexpected result in our setting, given the existence of a value-added system designed to counter such responses. Our results suggest caution when policymakers consider layering apparently unrelated accountability regimes on top of each other.

1 Introduction

The chronic under-performance of public schools in many countries has increasingly led policy makers to embrace reforms that hold educators more accountable. As a result, accountability incentives are now an integral feature of a host of modern-day public education systems, in North America and elsewhere. Their appeal stems from the promise of improving student achievement by raising the effort of teachers and schools, but at a cost likely to be far lower than reforms – notably class size reduction – that simply increase inputs (see Hoxby (2002)).

The most widespread accountability system in the US, associated with the federal No Child Left Behind Act of 2001 (henceforth 'NCLB'), sets performance targets based on schools' sociodemographic composition, as well as penalties for repeated target non-attainment. Being a proficiency scheme, NCLB creates incentives to focus on students at the margin of passing – a feature that has been well-documented in the literature (see Reback (2008), for example) – rather than students further away from the passing threshold. Such an emphasis may have undesirable consequences, not least leading to the neglect of students at the bottom of the distribution and thereby increasing inequality. Quite different incentives arise as part of the growth-oriented schemes already operating in several states.[1] These schemes set test score targets that condition on prior student scores, providing incentives for schools to focus on students throughout the entire achievement distribution. Under North Carolina's sophisticated 'ABCs' system, for instance, all teachers and the principal at a school receive a monetary bonus if the school meets expected

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

growth targets for its students as they progress from one grade to the next.[2]

In many states, federal and state accountability systems now operate concurrently, presenting teachers and schools with an especially complex incentive environment. In part, the complexity stems from the fact that the federal program was implemented on top of pre-existing reforms without considering the incentives already flowing from them. As is the case in North Carolina, educators were faced with starkly different incentives under many of the prior regimes than those that began to operate once NCLB took effect. From a public finance perspective, this prompts the question whether the *combined* effects of multiple apparently-unrelated accountability policies might be significantly different (and perhaps worse in an efficiency or equity sense) than those under a unified incentive system, devised by an idealized social planner.

Our primary goal is to document interactions between accountability systems that may lead to unforeseen or even unwanted effects, thereby shedding light on this potentially important design issue. At a general level, we would like to understand how different accountability reforms interact in influencing student outcomes; the North Carolina context is useful in this regard. Specifically, we ask whether the introduction of NCLB affected school incentives under the pre-existing ABCs, and if so, how?

Using rich longitudinal data covering all public school students in North Carolina, we first show aggregate evidence of a response to NCLB by documenting an impressive decline in the fraction of schools failing the ABCs in the year NCLB was introduced: the failure rate falls from 25 percent in 2002 to only 4 percent in 2003. We then document a dynamic interaction between the incentives under NCLB and those under the ABCs: since current test

[2]Macartney (2014) shows that the ABCs appear to engender a form of dynamic gaming, whereby schools have an incentive to focus more effort on students in later rather than in earlier grades.

scores become future targets under the ABCs growth formula, the strong response by schools to NCLB in 2003 caused a stark rise in ABCs targets in 2004, making it more difficult for teachers and school principals to receive monetary rewards.[3]

Prior research has argued, quite plausibly, that NCLB should have little effect (if any) in states with consequential accountability regimes already in operation (Dee and Jacob, 2011). Our results show not only that schools in North Carolina did respond quite strongly to NCLB's introduction, but that in doing so, they jeopardized their chances of satisfying the requirements under the ABCs. Exploiting the rich student-level data we have access to, we construct a measure of the ex-ante NCLB threat faced by each school in 2003, and show that even schools which had a very high likelihood of passing NCLB in 2003 responded to its introduction and made it more difficult for themselves to pass the ABCs in future years. This is a particularly interesting result, given that failing the ABCs means relinquishing a monetary bonus while the first failure of NCLB only places a school on a 'watch list.'

Despite evidence that schools with a relatively high probability of passing NCLB responded to its implementation, we show that schools were differentially affected. We use the rich longitudinal data and the fact that we observe a pre-NCLB period to predict the likelihood that each school would satisfy NCLB requirements in 2003 without making any changes to its teaching practices. Relative to schools with a high probability of satisfying NCLB, schools with a low probability of passing faced a greater NCLB threat and thus heightened incentives to respond. Consistent with this, we find that the 2003 improvements were most pronounced at schools in the lowest quar-

[3]Indeed, 2004 marks the first year in which the average school failed to satisfy its fifth grade ABCs math target.

tile in the distribution of the predicted NCLB passing probabilities. Using a difference-in-differences strategy, we show that students at these schools achieved (at the school-grade-level) half of a standard deviation higher test score gains – the metric used to evaluate schools under the ABCs – in the transition from third to fourth grade than students attending schools in the highest quartile.

Previous work on the effects of proficiency-count systems like NCLB provides evidence that educators typically respond by targeting resources toward the students for whom a marginal increase in the probability of reaching proficiency status would lead to the largest increase in the probability of the school having a successful outcome. Indeed, Neal and Schanzenbach (2010) show that NCLB caused the largest gains over predicted scores – the outcomes likely to prevail in the absence of NCLB – for fifth grade Chicago Public School students in the middle of the baseline achievement distribution. We relate our results to this previous work by documenting similar patterns for fourth grade math students in North Carolina. Specifically, we show that, relative to students with low and high probabilities of reaching proficiency status, students in the middle of the distribution realized the largest gains over their predicted math scores in 2003.

Given our research setting, this is a surprising result. Most researchers agree that value-added schemes like the ABCs are more sophisticated than proficiency-count systems because they provide incentives for schools to focus on students throughout the distribution rather than just a marginal subset of them. Yet we show that educators seemingly abandoned these more sophisticated incentives and opted to focus on students who were marginal under the new, inferior regime when it was introduced. These results shed new light

on the importance of accountability program interactions, illustrating how policymakers must exercise caution when combining distinct accountability incentives.

This descriptive evidence of a strong school response to the introduction of NCLB prompts the question *how* schools responded – through what channels? At heart, accountability reforms seek to raise the effort of teachers – a vital input in a school setting. In order to understand how outcomes are affected by accountability incentives, it is natural therefore to look at the mediating channel of teacher effort.

Building on the current analysis, we seek to shed light on how teacher and school effort change in response to accountability incentives. The measurement challenge in addressing this issue is obvious, given that teacher effort is typically unobserved. It calls for a strategy to isolate accountability-related factors that plausibly shift teacher effort. Here, we use two important aspects of the incentive environment: the differential incentive effects across schools arising from the introduction of the federal accountability program (just described), and the fact that a portion of the school's objective function can be written down explicitly – unusual in an education context – by using the provisions in the state and federal accountability schemes.

We argue that the introduction of NCLB in 2003 provides the relevant type of exogenous incentive variation needed to identify the effects of teacher effort. The new performance requirements under NCLB should affect effort decisions but not other determinants of student test scores – the reform's parameters do not appear in the production technology, nor do they determine any of its non-effort inputs.

This variation can be fed into a simple model that makes clear the way that

accountability incentives from more than one source affect student outcomes, focusing on teacher effort. Key here are the school objectives and the nature of the production technology. While the objectives of public schools are undeniably complex, a given accountability scheme will imply a known relationship between observed outcomes and stated rewards (or punishments). In the case of North Carolina's ABCs, for example, each school is presented with a target score that is a mechanical function of prior scores, based on a publicly-available formula; should the school's actual aggregate score in a given year exceed that target, then the school's teachers receive a pre-announced bonus payment. Similarly, under NCLB, each school faces proficiency targets for all students and nine subgroups.[4] The explicit nature of the reward formulae under each accountability scheme allows us to write down the school payoff from meeting performance targets, with teachers balancing the benefits of target attainment against the cost of effort. We pursue this line of research in related work.

The rest of the paper is organized as follows: the next section describes the institutional setting and the rich administrative data to which we have access. Section 3 presents preliminary evidence relating to interactions between NCLB and the ABCs at the aggregate level. In Section 4, we discuss the construction of a measure of ex-ante NCLB threat for each school, and in Section 5 we show that even schools that faced little threat responded to NCLB. Section 6 shows evidence of differential effects of the introduction of NCLB across schools facing varying levels of NCLB threat. In Section 7, we show that schools responded in part by directing additional resources toward students whom were on the margin of being proficient on standardized tests.

[4]Failure to attain these brings stigma; and repeated failure can result in draconian school takeovers and staff re-assignment.

We offer concluding remarks in Section 8.

2 Institutional Setting and Data

North Carolina provides a suitable context for our proposed study, for institutional and data reasons. On the institutional side, the state provides useful incentive variation arising under two separate accountability regimes. High-stakes accountability was implemented under the ABCs of Public Education legislation in the 1996-97 school year for all schools serving kindergarten through grade eight. Under the ABCs, each grade from three to eight in each school is assigned a school-grade specific gain target, which depends on both average previous student performance and on a constant level of expected growth. Based on average school-level gains across all grades in student standardized math and reading scores, the ABCs pays a monetary bonus to all teachers and the principal if a school achieves its overall growth target.

NCLB provisions were implemented in North Carolina in the 2002-03 school year following the passage of the federal No Child Left Behind Act in 2001. In contrast to the pre-existing pecuniary incentives under the ABCs, NCLB focuses on penalties for under-performing schools. The federal program aims to close performance gaps by requiring schools to satisfy Adequate Yearly Progress (AYP) targets for all students and for each of nine student subgroups. As a necessary condition for satisfying AYP, a school must ensure that the percentage of its students in each subgroup achieving proficiency status on state tests meets the state-mandated target. If a school fails to meet each of its subgroup-specific targets, it faces an array of penalties that become more severe over time in the event of repeated failure.

In addition to the institutional incentive variation, North Carolina has in-

credibly rich longitudinal education data from the entire state, provided by the North Carolina Education Research Data Center (NCERDC). These contain yearly standardized test scores for each student in grades three through eight and encrypted identifiers for students and teachers, as well as unencrypted school identifiers. Thus, students can be tracked longitudinally, and linked to a teacher and school in any given year. The demographic student-level information also allows us to track any effects of the accountability regimes for students belonging to different demographic groups.

Our sample period runs from 1997-2005. To focus on schools facing similar incentives, we limit the sample to schools serving kindergarten to eighth grade, and exclude special education, vocational, and alternative schools. We also only retain schools with a highest grade served between grades five and eight, thus avoiding the different accountability provisions that arise in high schools and the (potentially) different incentives in schools with only one or two high-stakes grades.

Table 1 provides sample summary statistics. Test scores are measured on a developmental scale, which is designed so that each additional point represents the same amount of knowledge gained, irrespective of the baseline score and the grade of testing. As mentioned, the ABCs program sets test score *growth* targets for schools, requiring that students demonstrate sufficient improvement as they progress through their educational careers. In contrast, NCLB holds schools accountable for test score *levels* by requiring that of each of nine student subgroups achieve proficiency status on standardized tests. The race, disability, limited English proficiency, and free lunch eligibility variables allow us to keep track of student subgroup memberships at each school in each sample year, thus enabling us to reconstruct school-level

Table 1: Summary Statistics

Student-Level			
	Mean	Standard Deviation	Number of Observations
Performance Measures			
Math Score			
Grade 3	144.67	10.67	905,907
Grade 4	153.66	9.78	891,969
Grade 5	159.84	9.38	888,467
Grade 6	166.43	11.12	892,087
Grade 7	171.61	10.87	884,286
Grade 8	174.76	11.63	860,623
Math Growth			
Grade 3	13.85	6.3	841,720
Grade 4	9.4	5.96	730,627
Grade 5	6.82	5.29	733,037
Grade 6	7.55	5.68	722,491
Grade 7	5.99	5.6	718,994
Grade 8	3.73	5.86	705,095
Reading Score			
Grade 3	147.03	9.33	901,233
Grade 4	150.65	9.18	887,147
Grade 5	155.79	8.11	883,685
Grade 6	156.79	8.85	889,445
Grade 7	160.3	8.19	882,288
Grade 8	162.79	7.89	859,089
Reading Growth			
Grade 3	8.15	6.72	837,361
Grade 4	3.75	5.55	725,590
Grade 5	5.61	5.21	727,864
Grade 6	1.54	4.95	718,291
Grade 7	3.77	4.92	716,496
Grade 8	2.76	4.62	703,236
Demographics			
College-Educated Parents	0.27	0.44	5,456,948
Male	0.51	0.5	5,505,796
Minority	0.39	0.49	5,502,665
Disabled	0.14	0.35	5,498,312
Limited English Proficient	0.03	0.16	5,505,479
Free or Reduced Price Lunch	0.42	0.49	3,947,605

School-Level			
	Mean	Standard Deviation	Number of Observations
Failed ABCs	0.27	0.45	14,052
Failed NCLB	0.37	0.48	5,014
Proficiency Rate	0.79	0.11	14,042

Notes: The sample excludes special education, vocational, and alternative schools. We also exclude high schools and schools with a highest grade served that is below grade five. Student-level summary statistics are calculated over all third to eighth grade student-year observations from 1997-2005 in the eligible schools. The free or reduced price lunch eligibility variable is not available prior to 1999. School-level summary statistics are calculated over all eligible school-year observations from 1997-2005. The NCLB performance indicator variable is not available prior to 2003.

outcomes from the underlying student-level data.[5] In the aggregate, about 39 percent of students are minorities, 14 percent are learning-disabled, only 3 percent are limited English-proficient, and 42 percent are eligible for free or reduced price lunch. With respect to the school-level variables, across the sample period, 27 percent of schools failed the ABCs and 37 percent failed NCLB. Recall that NCLB is a proficiency count system which assesses school performance according to the fraction of students achieving proficiency status on End-of-Grade tests. Throughout our sample period, the average school had a school-wide proficiency rate of 79 percent on math and reading tests.[6] We explore interesting time series variation in these rates next.

3 Preliminary Evidence: The Interactions Between the ABCs and NCLB

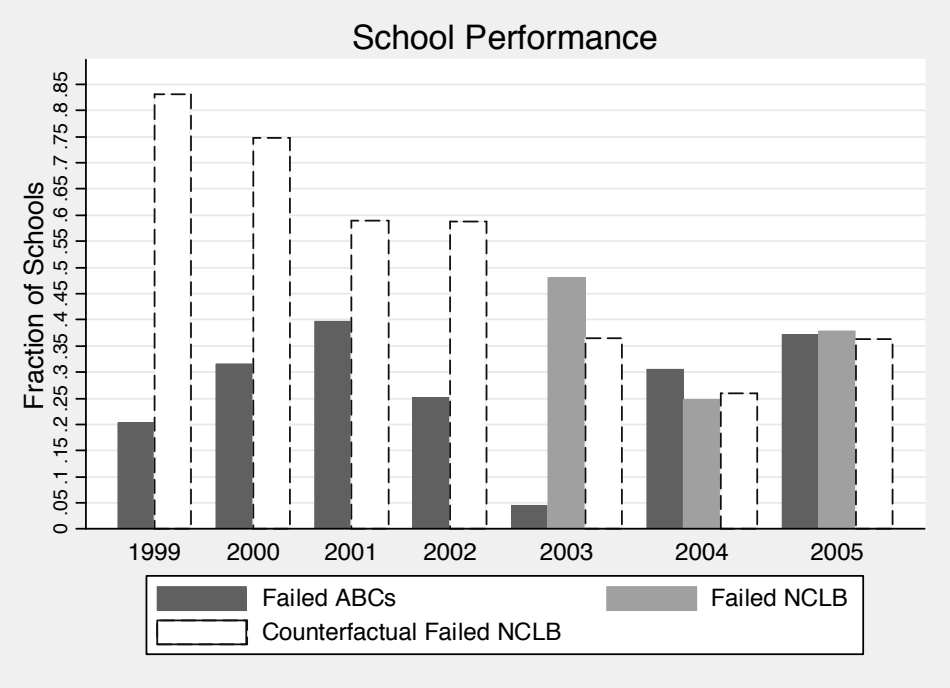
In North Carolina, NCLB was introduced into an environment in which schools were already accountable for performance under the ABCs (the state accountability program). To identify the effects of NCLB on student outcomes, Dee and Jacob (2011) use states that had pre-existing accountability programs as 'control' states, arguing, quite plausibly, that NCLB should have had little effect on school incentives here. We present evidence indicating that schools in North Carolina – a state with a pre-existing scheme – actually responded quite strongly to the introduction of NCLB, and that the responses affected future incentives under the ABCs.

This section focuses on aggregate evidence. We first show that an impres-

[5]The nine groups are white, black, Hispanic, Asian, Indian, multi-racial, disabled, free or reduced price lunch-eligible, and limited English proficient. In Table 1, to be concise, we report the fraction of minority (non-white) students.

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

sive overall improvement in school-level ABCs pass rates occurred in 2003, the year in which NCLB was implemented. We then document an interaction between the incentives stemming from the federal and state programs, graphically demonstrating the connection between NCLB-related improvements and ABCs targets. In particular, since current test scores become future targets under the ABCs, the response to NCLB in 2003 made it more difficult for schools to pass the ABCs in 2004.



Notes: The figure shows (a) the fraction of schools failing the ABCs (all years), (b) the fraction of schools failing NCLB (2003-), and (c) the fraction of schools predicted to fail NCLB on a consistent basis from calculations using the underlying student-level data (all years).

Figure 1: School Performance from 1999 to 2005

Figure 1 shows the fraction of schools failing the ABCs and NCLB in each year, starting in 1999. We construct a consistent series showing the counterfactual NCLB failure rate in the years prior to 2003 by applying the NCLB outcomes calculations used in 2003 to the underlying student-level performance data. Specifically, we first use the student-level demographic information to construct NCLB subgroup memberships prior to NCLB at

each school. Following the NCLB provisions that operate in North Carolina, we hold a school accountable for the performance of a subgroup of students only if that subgroup's membership is at least 40 students in a given year. We then use the NCLB test score and proficiency rate targets that prevailed in 2003 to determine whether all of the school's accountable subgroups satisfied AYP targets,[7] and assign the school a counterfactual NCLB 'pass' or 'fail' status in each year based on this information.

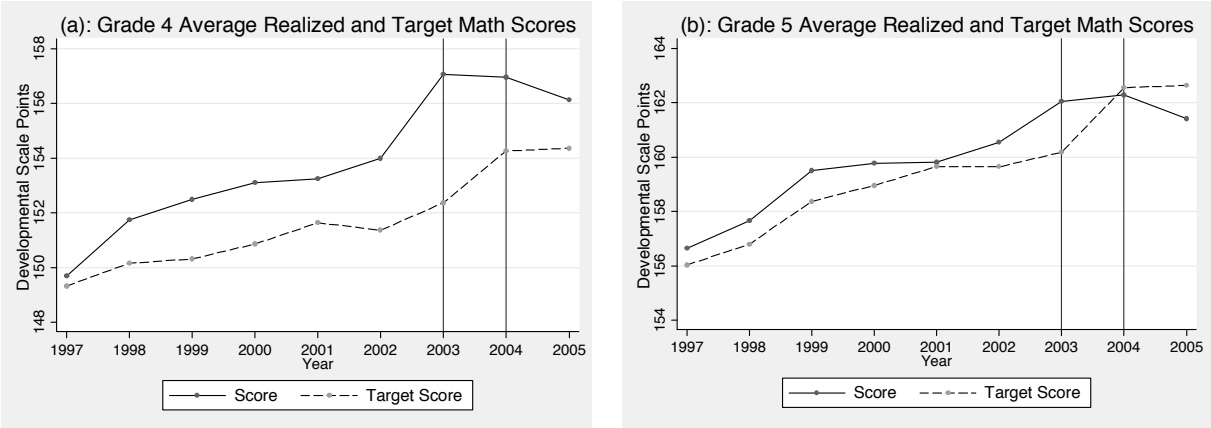
The year in which NCLB was introduced, 2003, is associated with a remarkable decline in the fraction of schools failing the ABCs, down from 25 percent in 2002 to only 4 percent in 2003. To lend credence to the notion that this reflects an NCLB-triggered response, notice that the fraction of schools predicted to fail NCLB declines substantially from 58 percent in 2002 to 36 percent in 2003, consistent with schools taking steps to improve along the dimensions required under NCLB.[8] The figure also shows that the ABCs improvement was short-lived, as the failure rate quickly jumps back, to over 30 percent in 2004.

The quick reversion of the ABCs failure rate in 2004 reflects a (perhaps unintended) interaction between the incentives under NCLB and those under the ABCs. Being a growth-oriented system, the ABCs encourages educators to think about dynamic considerations, as test scores in year t become targets in year $t+1$. In contrast, under NCLB, educators operate in a static incentive environment in which the chief concerns are current proficiency rates. As we show next, while each incentive system may work well in isolation, the inter-

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

[8]Due to the many nuances associated with the implementation of NCLB over the years, we are unable to perfectly reproduce school-level outcomes in the post-NCLB period, though we are able to do a reasonable job. Other researchers using the NCERDC data have faced similar difficulties (see, for example, Ahn and Vigdor, 2013).

action of the two provides educators with potentially conflicting incentives.



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

Figure 2: ABCs and NCLB Interaction

Figure 2 outlines how NCLB's introduction disrupted schools' ABCs incentives and, ultimately, their ABCs outcomes. The figure reports average test scores and average ABCs score targets for math. The vertical lines depict the year in which NCLB was implemented, 2003, and the year immediately after, 2004. The figure shows that, relative to other years, average math scores in both fourth and fifth grade display impressive improvements from 2002 to 2003, the solid lines rising quite sharply. These improvements led to the large decline in the fraction of schools failing the ABCs shown in Figure 1.

Observing the trends around 2004, however, it is apparent that schools responded aggressively to NCLB, as the 2003 improvement made it more challenging for them to reach their ABCs targets in subsequent periods: by raising test scores in response to NCLB, schools improved contemporaneous ABCs performance but simultaneously increased future ABCs targets and, correspondingly, lowered the likelihood of future ABCs success (as indicated by the steeply rising dashed lines in both panels). The dynamic interac-

tion between the two schemes is seen clearly by observing the pronounced increase in the average fifth grade target in 2004, a direct consequence of the pronounced increase in the average fourth grade score in 2003. In fact, 2004 marks the first year in which the average fifth grade target rose above average performance, implying that the average school failed to satisfy its fifth grade math requirement. Thus, by improving students' performance in 2003 so drastically, many schools set baseline scores that they were unable to improve upon sufficiently in 2004.

In the next section, we describe how we construct a measure of NCLB threat for each school. With this measure in hand, in Section 4, we consider schools that regularly passed the ABCs prior to 2003 and that had a very high likelihood of satisfying NCLB requirements, showing that even some of them responded to NCLB and correspondingly made it more difficult for themselves to attain the monetary bonuses associated with the ABCs in 2004.

4 NCLB Incentive Strength Measures

Having presented motivating evidence for the interactions between the ABCs and NCLB, we now describe the construction of measures that allow us to identify differential responses to NCLB across students and schools. We use these measures to group schools according the strength of the NCLB-related incentives they face and to investigate whether the strongest responses to NCLB are observed in the schools that are most strongly affected by NCLB's introduction.

Constructing Student-Level Proficiency Probabilities: We first describe how we construct a measure capturing a school administrator's predication of the probability that each student will reach proficiency status on

EOG math and reading tests. The measure is constructed in three steps. Suppose the goal is to estimate the likelihood that students in grade g and year t^* will reach proficiency status on the EOG test in subject j. First, we use the entire (state-wide) sample of students in grade g in all years up to $t^* - 1$ and estimate the following logit model for each subject-grade pair separately:

$$P_{ijgt} = \begin{cases} 1 & \text{if } \mathbf{x}'_{igt}\beta - \epsilon_{ijgt} > 0 \\ 0 & \text{otherwise} \end{cases} \quad \forall\ t \leq t^* - 1, \tag{1}$$

where P_{ijgt} is an indicator for student i in grade g and year t having reached proficiency status on the EOG test for subject j. The vector $\mathbf{x}'_{igt}$ includes quadratics in prior year math and reading EOG test scores and demographic indicators for parental eduction, ethnicity, gender, limited English proficiency status, free or reduced price lunch eligibility, and disability status. Second, we retain the estimated coefficients, $\hat{\beta}_{jg}^{t^*-1}$. Third, using the logistic distribution, the estimated coefficients, $\hat{\beta}_{jg}^{t^*-1}$, and the observable characteristics of the students in grade g in year t^*, $\mathbf{x}'_{igt^*}$, we predict the probability that each student i in grade g and year t^* will achieve proficiency status on the EOG test for subject j, $\hat{P}_{ijgt^*}$.

These predictions represent school administrators' assessments of the likelihood that each student will reach proficiency status. Intuitively, schools use the relationship between student observable characteristics and test score outcomes in previous years to predict the likelihood that each student will achieve proficiency status in the current year. Since $\hat{P}_{ijgt^*}$ is formed using the relationship between student characteristics and outcomes in past years, when $t^* = 2003$, it represents the probability that student i in grade g will

achieve proficiency status if the school makes no NCLB-related adjustments – that is, if no additional resources are directed toward that student.

Constructing Subgroup-Level Passing Probabilities: To calculate the probability that a subgroup passes NCLB, we aggregate the predicted proficiency probabilities of the students who are members of each subgroup, and use the standard normal distribution to determine the probability that the subgroup reaches proficiency status. To illustrate, consider a given subgroup, R_1, in school s and time t. Suppose there are $N_{st}^{R_1}$ student members of this subgroup and denote each student's test score at the end of year t as y_{isgt}. In terms of NCLB targets, let y_{gt}^* be the grade-specific test score proficiency target and let p_t^* denote the percentage of students within a subgroup who must be proficient in order for the subgroup to meet AYP. The probability that subgroup R_1 meets AYP at the end of year t may be written as

$$\Pr\left(\sum_{i\in R_1}\frac{1(y_{isgt}\geq y_{gt}^*)}{N_{st}^{R_1}}\geq p_t^*\right) \tag{2}$$

The proficiency rate of subgroup R_1 is the mean of $N_{st}^{R_l}$ bernoulli random variables, each of which take on a value of 1 when student i reaches proficiency and zero otherwise. From above, let $\hat{P}_{ijgt}$ be the probability that student i reaches proficiency. We make the simplifying assumption that unexpected changes to student test scores are uncorrelated within a subgroup. In this case, the expected pass rate in subgroup R_1 is $\mathrm{E}(R_1)=\frac{\sum_{i\in R_l}\hat{P}_{ijgt}}{N_{st}^{R_l}}$ and the standard deviation of the pass rate is given by $\mathrm{S}(R_1)=\frac{\sqrt{\sum_{i\in R_l}\hat{P}_{ijgt}\left(1-\hat{P}_{ijgt}\right)}}{N_{st}^{R_l}}$. The proficency rate for subgroup R_1 is an average of $N_{st}^{R_1}$ student indicator variables, implying that it is normally distributed and that the probability of

the subgroup passing NCLB can be written as

$$\Phi\left(\frac{\mathrm{E}(R_1) - p_t^*}{\mathrm{S}(R_1)}\right) \tag{3}$$

where denotes $\Phi(\cdot)$ is the CDF of the standard normal distribution.

Constructing School-Level Probabilities of passing NCLB: We define our measure of school-level NCLB threat or incentive strength as school administrators' prediction of the likelihood that the school will pass NCLB. To form this prediction, we make the simplifying assumption that the subgroup-specific probabilities of reaching proficiency are independent within a school. We therefore take all of the numerically significant subgroups within each school and multiply the subgroup-subject-specific probabilities together (given by equation (3)) to arrive at a school-level probability of passing NCLB.

Table 2 provides summary statistics for our measures of NCLB incentive strength. In Panel A, we compare the realized subject-specific proficiency rates and the rates we predict using equation (1). In each year, we do a fairly good job of predicting the actual subject-specific pass rates that occur. Interestingly, the largest deviation between realized and predicted rates occurs in 2003. This is consistent with schools making adjustments in 2003, as our predictions from that year are based solely on the relationship between student observable characteristics and outcomes that prevailed prior to NCLB.

Our measure of NCLB incentive strength is the ex-ante probability that a school will pass NCLB if it makes no changes to its teaching methods. We calculated this measure for each school-year pair from 2003 to 2005. In Panel B of Table 2, we group schools in each year into quartiles of the school-level distribution of NCLB incentive strength. While schools in the first and second

Table 2: NCLB Measures Summary

Year	Math Rate	Predicted Math Rate	Reading Rate	Predicted Reading Rate
		Panel A: Student Proficiency Rates		
2000	0.79	0.82	0.75	0.78
2001	0.82	0.82	0.77	0.79
2002	0.85	0.84	0.79	0.8
2003	0.89	0.86	0.85	0.81
2004	0.89	0.9	0.85	0.86
2005	0.88	0.91	0.86	0.86

Year	Quartile 1	Quartile 2	Quartile 3	Quartile 4
		Panel B: School-Level Probability of Passing NCLB		
2003				
Passed NCLB	0.25	0.34	0.59	0.85
Predicted to Pass NCLB	0	0.013	0.45	0.98
N	312	444	480	401
2004				
Passed NCLB	0.19	0.44	0.8	0.95
Predicted to Pass NCLB	0	0.013	0.49	0.98
N	166	272	427	793
2005				
Passed NCLB	0.14	0.45	0.83	0.94
Predicted to Pass NCLB	0	0.013	0.48	0.98
N	358	403	468	442

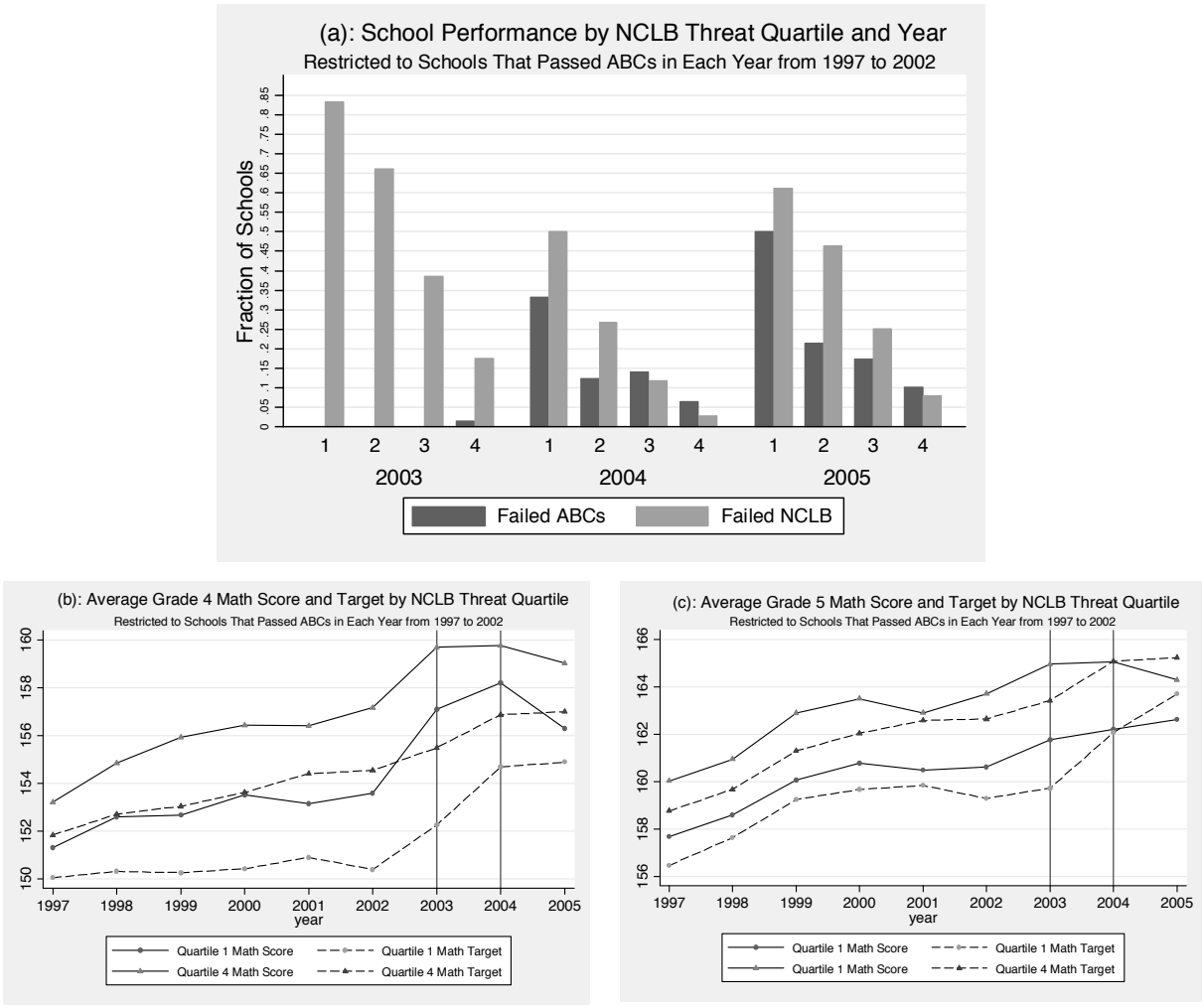
Notes: The sample excludes special education, vocational, and alternative schools. We also exclude high schools and schools with a highest grade served that is below grade five. Student-level summary statistics are calculated over all third to eighth grade student-year observations from 2000-2005 in the eligible schools. School-level summary statistics are calculated over all eligible school-year observations from 2003-2005.

quartiles have very low predicted probabilities of passing, our measure does predict actual school-level NCLB outcomes. For example, in 2003, only 25 percent of schools in the first quartile passed NCLB compared to 85 percent of schools in the fourth quartile. Thus, the schools for which we predict the highest ex-ante probability of passing NCLB are indeed the schools that exhibit the highest realized pass rate.

5 NCLB's Distortion to ABCs at the Top Schools

In this section, we show that even schools that had a high probability of passing NCLB responded to its implementation and that this response potentially jeopardized their future ABCs outcomes. We begin by calculating for each school its quartile position in the 2003 NCLB incentive strength distribution.

We then restrict the sample to schools that are present in the data for all years from 1997 to 2005, and, in each of these years, we assign the school its *2003 NCLB incentive strength quartile*. In Figure 3 below, we then consider only schools that passed the ABCs in each year from 1997 to 2002. These are the schools that were consistently earning ABCs bonuses for their teachers.



Notes: The sample is restricted to schools that appear in all sample years and that passed the ABCs in each year from 1997 to 2002. We divide schools into quartiles of the 2003 distribution of the probability of passing NCLB. Panel (a) plots the ABCs and NCLB failure rates across schools within each quartile. In panel (b), we present trends in average grade four math scores and grade four math targets for schools in the first and fourth quartile. In panel (c), we present trends in average grade five math scores and grade five math targets for schools in the first and fourth quartile.

Figure 3: Dynamic Distortion of ABCs

In Figure 3(a), we plot the ABCs and NCLB failure rates for these schools, showing the respective rates within each of the quartiles of NCLB incentive

strength. There are 18 schools in the first quartile, 56 in the second quartile, 127 in the third quartile, and 137 in the fourth. The ABCs failure rates in each of the quartiles start to rise after 2003. This is the case even among schools in the fourth quartile, those that had the highest probability of passing NCLB (recall that 85 percent of these schools actually passed in 2003). By 2005, 10 percent of these schools are failing the ABCs.

Figures 3(b) and 3(c) present evidence that is consistent with these schools making it more difficult for themselves to pass the ABCs in 2004 and 2005 as a result of their actions in 2003. In particular, we present similar evidence to Figure 2 but for schools in the first and fourth quartiles of the 2003 NCLB incentive strength distribution only. Both types of school display similar patterns in their NCLB responses: they improve grade four performance in 2003 but this improvement leads to higher grade five targets in 2004. Indeed, even among the schools that had the highest likelihood of passing NCLB, the average fifth grade math target exceeds the average fifth grade math score for the first time in 2004 and continues to exceed it in 2005.

Having presented evidence consistent with schools that were regularly passing the ABCs being disrupted by NCLB's introduction, we next explore differential responses to NCLB across schools, assessing where the NCLB-related gains in student performance were highest.

6 Differential Effects

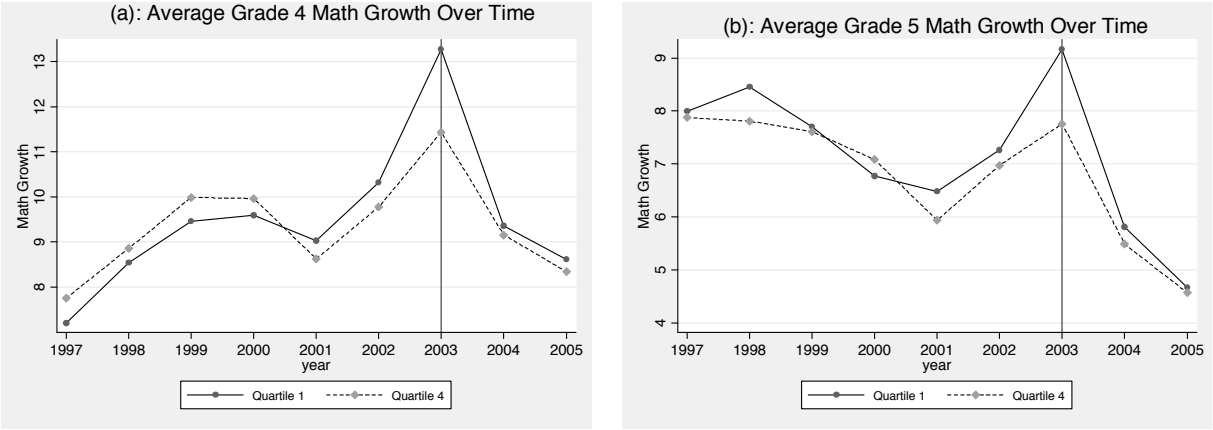
We now explore whether the largest responses to NCLB in 2003 are seen among the schools that faced the highest level of NCLB threat. We first categorize each school according to its position in the school-level distribution of the predicted probability of passing NCLB in 2003. Each school is assigned

its position in the 2003 distribution for all years from 1997 to 2005. To assess whether schools that were more in danger of failing NCLB responded more, we compare average grade-level math gains in 2003 between schools in the fourth quartile (lowest-threat schools) and schools in the first quartile (highest-threat schools) of the 2003 distribution. If NCLB provides stronger incentives for schools that are more likely to fail it, we would expect higher gains in 2003 at schools in the first quartile relative to schools in the fourth quartile. We ensure that we track the same schools over time by *restricting the sample to include only the schools that appear in the data in each year from 1997 to 2005.* To address mean reversion and other threats to identification, we implement a difference-in-differences (DiD) strategy, comparing the difference in gains between low- and high-threat schools in 2003 to the difference in gains in 2002. Specifically, we construct the following DiD estimator:

$$\underbrace{(y_{2003}^{q1} - y_{2003}^{q4})}_{\text{Difference in 2003}} - \underbrace{(y_{2002}^{q1} - y_{2002}^{q4})}_{\text{Difference in 2002}} \tag{4}$$

which compares the difference between the average first and fourth quartile gain scores after NCLB to the same difference in 2002. Since the NCLB effect is present only in 2003, by comparing 2003 differences to 2002 differences, we eliminate the mean reversion threat and isolate the effect of NCLB's introduction.

Table 3 presents the DiD analyses while Figure 4 provides the complementary visual evidence. To be clear, the vertical axes in Figure 4 measure the degree to which annual average student scores improve from grade $g-1$ to grade g, and school-grade observations are categorized according to whether they were in the first or fourth quartile of the 2003 probability of passing



Notes: The sample is restricted to schools that appear in all sample years. We divide schools into quartiles of the 2003 distribution of the probability of passing NCLB. We then present trends in average grade four math growth scores in panel (a) and trends in average grade five math growth scores in panel (b).

Figure 4: Differential Effects of NCLB

NCLB distribution.

Table 3: Difference-in-Differences Estimates of Differential NCLB Effects

	Grade 4			Grade 5		
	(1)	(2)	(3)	(4)	(5)	(6)
	Post	Pre	Difference:	Post	Pre	Difference:
School Rank	2003	2002	Post-Pre	2003	2002	Post-Pre
First Quartile	13.27***	10.31***	-	9.17***	7.27***	-
	(0.247)	(0.195)	-	(0.16)	(0.153)	-
Fourth Quartile	11.43***	9.77***	-	7.76***	6.98***	-
	(0.151)	(0.136)	-	(0.095)	(0.118)	-
Difference	1.84***	0.54**	1.3***	1.41***	0.29	1.12***
(First − Fourth)	(0.29)	(0.237)	(0.268)	(0.186)	(0.193)	(0.213)

Notes: Columns (1) and (4) show the average grade-level math growth score across all schools in 2003. Columns (2) and (5) show the average grade-level math score across all schools in 2002. Column (3) shows the difference between the third row entries of Column (1) and Column (2). Column (6) shows the difference between the third row entries of Column (4) and Column (5). 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.

We first consider the effects of NCLB's introduction in fourth grade. The left panel of Table 4 presents the DiD estimates and Figure 4(a) presents the associated time trends. In 2002, the difference between the average gains of low- and high-threat schools was 0.54 developmental scale points, or 0.23 (grade-level) standard deviations. In transition from third to fourth grade in

2003, however, students at schools that faced high NCLB threat gained 1.84 developmental scale points worth of knowledge more than students who were at schools that faced low threat. The DiD estimate comparing 2003 and 2002 shows that, relative to schools in the fourth quartile, NCLB's introduction caused a 1.3 point higher average gain score at schools in the first quartile. In terms of the 2002 fourth grade (grade-level) distribution in math gains, NCLB is thus associated with 0.55 standard deviation higher gain scores at schools that were least likely to pass NCLB. The same estimate for fifth grade, reported in the right panel in Table 3, is 1.12 scale points, a magnitude of 0.63 standard deviations. Figure 4(b) provides the complementary visual evidence for fifth grade.

7 Student-Level Resource Reallocation

Since NCLB is a proficiency-count scheme, when deciding how to target resources toward students in response to NCLB, a school will form a prediction over how likely it is that a student will achieve proficiency status on EOG tests. The school will also aggregate these student-level probabilities to form predictions over the likelihood that it will satisfy all of its subgroup-specific proficiency rate targets, and it will then arrive at final prediction over the likelihood that the entire school passes NCLB. The level of 'extra' resources devoted to each student will depend on both the probability that the school satisfies NCLB and the probability that an individual student reaches proficiency status on standardized tests. Holding constant the school-level probability, the most cost effective strategy for a school is to direct resources toward the students for whom a marginal increase in the probability of achieving proficiency status would lead to the largest increase in the probability that the

school passes NCLB (Reback, 2007). These are typically the students who have neither a very low nor very high probability of reaching proficiency status without additional resources being directed toward them. Students in the tails of the probability distribution either require a prohibitively costly amount of additional effort to meaningfully alter the probability or will almost surely pass without any additional attention; in either case, a proficiency scheme like NCLB creates incentives for schools to direct resources away from the tails and toward the middle of the distribution.

If one were thus to plot NCLB-related effort against student-level proficiency probabilities, one would expect to find an inverted-U relationship: additional effort should be low for students who have a low probability of passing, high for students who are on the margin of passing, and low for students who have a high likelihood of passing without additional teaching resources. This is the central idea in Neal and Schanzenbach (2010), who show that NCLB caused the largest gains over predicted test scores – the outcomes that would have prevailed in the absence of NCLB – for fifth grade students in the middle of the baseline achievement distribution. We explore whether similar trends prevail in North Carolina by using deviations from predicted test scores to infer teaching effort.

For each student, we predict the test score that would have prevailed in 2003 had NCLB not been implemented. Similar to the construction of the predicted proficiency probabilities, when predicting test scores for students in year t^*, we run the following regression separately for each subject-grade pair using data on all students up to year $t^* - 1$:

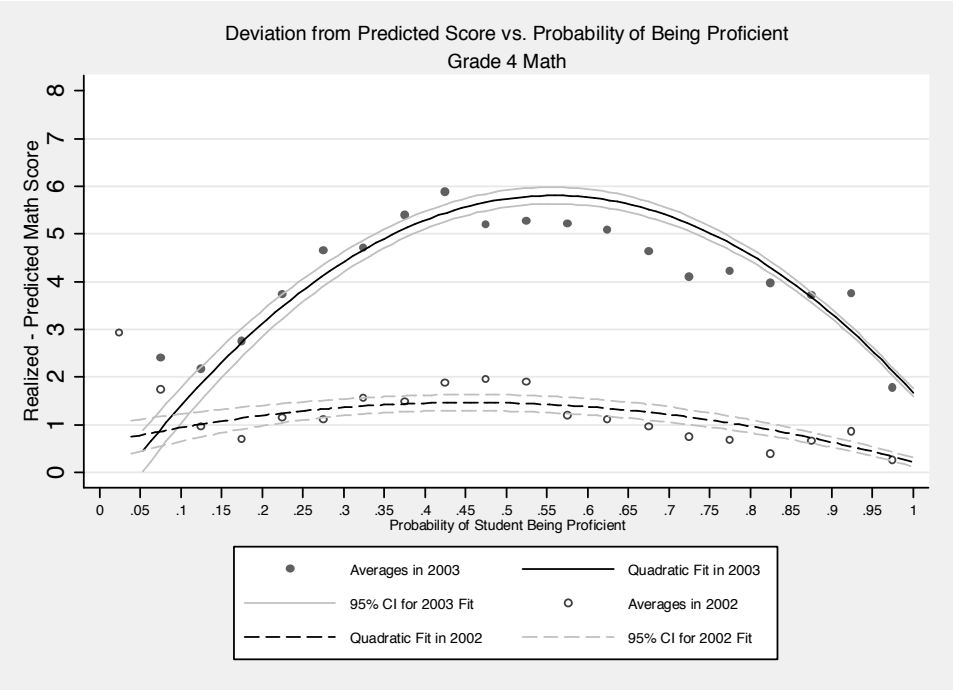
$$y_{ijgst} = \mathbf{x}'_{igt}\gamma \ + \nu_{ijgt}, \tag{5}$$

where the EOG test score in subject j for student i in grade g and year t is regressed on the same covariates as in equation (1). To predict test scores for students in year t^*, we use the estimated coefficients, $\hat{\gamma}_{jg}^{t^*-1}$, and the observable characteristics of students in year t^*, $\mathbf{x}'_{igt^*}$, to form the linear prediction, $\hat{y}_{ijgst^*} = \mathbf{x}'_{igt^*}\hat{\gamma}_{jg}^{t^*-1}$. When $t^* = 2003$, these predictions represent the typical test score that would prevail for a student with observable characteristics, $\mathbf{x}'_{igt^*}$, in the absence of any NCLB-related effort responses. To assess which students realized the most stark deviations from these typical scores, we calculate the difference between student i's realized score and his or her predicted score, $e_{ijgst^*} = y_{ijgst^*} - \hat{y}_{ijgst^*}$, and then plot these deviations against student i's predicted probability of reaching proficiency, $\hat{P}_{ijgt^*}$. If schools direct direct resources away from students with very low and very high probabilities of passing, and toward those who are marginal, one would expect to find an inverted-U relationship between e_{ijgst^*} and $\hat{P}_{ijgt^*}$.

Figure 5 shows the estimated quadratic relationship between student-level deviations of realized and predicted fourth grade math scores and proficiency probabilities. As expected, the relationship in 2003 follows an inverted-U pattern: relative to students in the middle of the distribution, students with low and high probabilities of reaching proficiency status realize test scores closer to the outcomes that would have prevailed had NCLB not been implemented. For example, the estimated quadratic relationship suggests that the average student with a proficiency probability of 0.5 is predicted to score 5.74 scale points higher than the test score that would have prevailed in the absence of NCLB. The same predictions for students with probabilities of 0.1 and 0.9 are 1.39 and 3.33 scale points, respectively. To put these differences into perspective, student-level deviations of realized and predicted scores have a

standard deviation of 4.58 scale points in 2003, implying that students with a 50 percent chance of being proficient exceeded their expected scores by 0.95 standard deviations more than students with a 10 percent chance and by 0.53 standard deviations more than students with a 90 percent chance.

Of note, the placebo estimates in 2002 do not reveal the same type of inverted-U pattern, as one would expect, given that NCLB had yet to be implemented. Specifically, for students with proficiency probabilities of 0.1, 0.5, and 0.9 in 2002, the predicted deviations between realized and expected scores are 0.93, 1.45, and 0.63 scale points, respectively. It is also worth noting the deviations from predicted scores in 2002 are much closer to zero at each point of the proficiency probability distribution, consistent with the idea that our measure of predicted test scores coincides well with realized test scores prior to NCLB and is therefore a good proxy for the outcomes that would have prevailed had NCLB not been implemented.



Notes: The figure is a binned scatter plot in which the difference between students' realized and predicted math scores is plotted against the probability that a student will reach proficiency status on the End-of-Grade four math test. To construct the figure, we first group students into 5 percentage point width bins according to the their proficiency probabilities. Within each bin, we then calculate the average of the difference between students' realized and predicted math scores. The circles represent these bin-specific averages. The solid circles depict the year NCLB was introduced, 2003, and the hollow circles represent the year prior, 2002. The dark solid line represents the quadratic relationship between the outcome of interest and the proficiency probabilities estimated on the underlying student-level data in 2003; the dashed line represents the same relationship in 2002. The figure reports the 95 percent confidence intervals corresponding to the predictions from the underlying regression specifications. Standard errors are clustered at the school level.

Figure 5: The Inverted-U in Grade 4 Math

8 Conclusion

Throughout the past decade, several states have operated both state-level and federal accountability programs concurrently, often with disregard for the potentially conflicting incentives that can arise between the two regimes. Using school administrative data, we have shown that even a sophisticated value-added scheme like the ABCs can be undermined by the introduction of a decidedly less efficient proficiency-count system. Schools in North Carolina responded aggressively to the introduction of NCLB, jeopardizing their chances of winning bonus payments under the state's scheme and redirecting resources across students in ways that value-added schemes are specifically designed to prevent.

In future work, we intend to use sensible modeling assumptions and the reduced-form interaction we have identified to determine the nature of education production within schools. In particular, we aim to make use of the variation in accountability incentives to uncover changes in teacher and school effort, a nontrivial feat given such quantities are typically unobserved.[9] With this in hand, we then plan to investigate the extent to which such effort contributes to student learning through test scores, according to both its persistence and interaction with other inputs in the education production technology.

[9]Focusing on effort is natural, given that accountability reforms are designed with raising it in mind.

References

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

[2] Burgess, Simon, Carol Propper, Helen Slater, and Deborah Wilson. 2005. "Who Wins and Who Loses from School Accountability? The Distribution of Educational Gain in English Secondary Schools." CEPR Discussion Paper No. 5248, September.

[3] Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2013. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." Working Paper 19423. NBER, Cambridge, MA.

[4] Cullen, Julie and Randall Reback. 2006. "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System," in *Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics*, edited by T. Gronberg and D. Jansen. Volume 14. Amsterdam: Elsevier Science.

[5] Dee, Thomas S. and Brian Jacob. 2011. "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management.* 30(3): 418-446.

[6] Deming, David J., Sarah Cohodes, Jennifer Jennings, Christopher Jencks, and Maya Lopuch. 2013. "School Accountability, Postsecondary Attainment and Earnings." Working Paper 19444. NBER, Cambridge, MA.

[7] Figlio, David N. and Joshua Winicki. 2005. "Food for thought? The effects of school accountability plans on school nutrition." *Journal of Public Economics*. 89(2-3): 381-94.

[8] Hoxby, Caroline M. 2002. "The Cost of Accountability." Working Paper 8855. NBER, Cambridge, MA.

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

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

[11] 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.

[12] Macartney, Hugh. 2014. "The Dynamic Effects of Educational Accountability." NBER Working Paper 19915. NBER, Cambridge, MA.

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

[14] 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.

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

[16] Reback, Randall, Jonah Rockoff, and Heather L. Schwartz. 2011. "Under Pressure: Job Security, Resource Allocation, and Productivity in Schools Under NCLB." Working Paper 16745. NBER, Cambridge, MA.