The Dynamic Effects of Educational Accountability

Hugh Macartney, *Duke University and National Bureau of Economic Research*

This paper provides the first evidence that value-added education accountability schemes induce dynamic distortions. Extending earlier dynamic moral hazard models, I propose a new test for ratchet effects, showing that classroom inputs are distorted less when schools face a shorter horizon over which they can influence student performance. I then exploit grade span variation using rich educational data to credibly identify the extent of dynamic gaming, finding compelling evidence of ratchet effects based on a triple-differences approach. Further analysis indicates that these effects are driven primarily by effort distortions, with teacher reallocations playing a secondary role.

I. Introduction

Against a backdrop of chronic underperformance in education, policymakers have increasingly embraced reforms that hold educators more accountable for the academic performance of their students. Such accountability measures have included standardized testing, publishing results that are comparable across schools and, more recently, providing high-powered incentives for both teachers and schools by awarding bonus pay if test scores exceed a specified target.

The way accountability targets are constructed is of particular interest from an incentive design perspective. Simple proficiency-based schemes, such as the one used under the 2001 federal No Child Left Behind Act, set performance targets that are independent of student, teacher or school measures past or present. The problem with such schemes is well known: they incentivize schools to focus on marginal students at the expense of non-

I would like to thank Robert McMillan, Aloysius Siow and Carlos Serrano for their guidance and support throughout this project. Thanks also to Gustavo Bobonis, Branko Boskovic, Raj Chetty, Damon Clark, Stephen Coate, Elizabeth Dhuey, Amy Finkelstein, Kirabo Jackson, Sacha Kapoor, Steven Lehrer, Joshua Lewis, Parag Pathak, Uros Petronijevic, Petra Todd, Trevor Tombe, Jacob Vigdor, and conference and seminar participants for their helpful suggestions. Remote access to the data for this study was generously provided by the North Carolina Education Research Data Center (NCERDC). I gratefully acknowledge financial support from the CLSRN Fellowship and the Royal Bank Graduate Fellowship in Public and Economic Policy. All remaining errors are my own. Information concerning access to the data used in this article is available as supplementary material online. Contact the author at hugh.macartney@duke.edu.

marginal ones.[1] In contrast, more refined value-added schemes provide incentives to focus on students throughout the distribution, by conditioning targets on prior scores to adjust for heterogeneity in education inputs. As a result of this desirable feature, such sophisticated schemes have become increasingly popular, having been implemented in Arizona, Colorado, Florida, North Carolina, South Carolina and Texas, among others.[2]

This paper is the first to draw attention to an important potential dynamic distortion arising from these more refined schemes. In particular, targets that depend on lagged achievement become manipulable with time, as raising effort under such a scheme not only affects the likelihood of exceeding the current target but also determines the target that follows. Given the implication that a strong performance today makes it more difficult to reap a bonus tomorrow, agents (teachers in this application) may become less responsive to the reform than they would be in the absence of dynamic considerations – an instance of the so-called 'ratchet effect.'

To shed light on the extent to which these dynamic distortions matter in practice, I first extend prior ratchet effect models to a finite-horizon setting and allow the choice variable to determine both contemporaneous and future output (in addition to targets).[3] The classic theory work in the area, notably Weitzman's seminal 1980 paper, features workers who make effort choices facing an infinite horizon, where targets depend on earlier output. This previous work yields the intuitive prediction that agents should identically suppress effort in every period, yet it is not amenable to empirical testing as the same pattern can emerge from a rival mechanism unrelated to, but potentially coexisting with, ratcheting behavior.

To clarify this point, the rationale behind conditioning on prior output is to adjust for heterogeneity in inputs, so that worker effort matters at the margin. The ideal (and fully efficient) target, which elicits the maximal effort for a given bonus level, would condition on

[1]Instances of gaming at the margin include redirecting resources from untested to tested subjects (Ladd and Zelli 2002), exempting disadvantaged students from testing (Cullen and Reback 2006), 'teaching to the distribution' of students (Neal and Schanzenbach 2010), and overt cheating (Jacob and Levitt 2003).

[2]Both types of high-powered schemes have tended to result in improved educational outcomes, as evidenced by research evaluating both broad definitions of accountability and specific performance-contingent systems (see Carnoy and Loeb 2002; Lavy 2002, 2009; Hanushek and Raymond 2005; Figlio and Kenny 2007; Dee and Jacob 2011; Muralidharan and Sundararaman 2011).

[3]The insight from modeling the production technology is that the degree of dynamic gaming depends on the extent to which the target coefficient deviates from the growth in output over time.

all factors beyond a worker's control so as to focus on pure effort. However, practical issues, such as ensuring agents understand the scheme enough to respond to it and incomplete information, are likely to force the policymakers to select an imperfect target. In particular, they might intentionally neglect some available prior information for transparency reasons[4] or be faced with some pertinent inputs that do not have observable proxies. If these imperfections occur systematically over time, then worker effort would be lowered across periods in a way that cannot be distinguished from the infinite horizon ratchet prediction.

Recasting the dynamic incentive problem in a finite-horizon context yields a new test for ratchet effects that is clearly distinguishable from responses to such systematic imperfections. The model I set out involves a single effort-making body, setting effort in light of the prevailing incentives. It is intended to capture key aspects of the particular incentive scheme I consider: the North Carolina accountability system established in 1996. Under that scheme, all teachers and the principal at a school receive a monetary bonus if the school meets specified growth targets in student achievement, where those targets condition on the average prior test scores of students. This dependence implies that students contribute to the school aggregate target only as long as they remain in the school, thereby determining the finite horizon faced by the school principal, which in turn affects the extent of the dynamic gaming. As the horizon becomes shorter, the downside associated with high performance is mitigated since there are fewer future periods in which the target will be raised, and so teacher effort will tend to be increased.

The theory lends itself naturally to empirical testing, given that the principal's horizon can be captured by the grade span of the school. The fact that I observe multiple grade-span configurations (in particular, K-5, K-6 and K-8) in North Carolina suggests a viable and transparent identification strategy: Comparing teacher behavior *in a particular grade* across schools with different grade spans, the model implies that schools serving fewer future grades should exert greater effort than those serving a greater number of future grades. For example, grade five teachers at K-5 schools are predicted to exert a higher level of effort than their K-8 or K-6 counterparts, since the negative externality that a K-5 school imposes on a 6-8

[4]For example, they might only condition on the prior output of one agent, rather than all previous measures for all relevant agents.

school through high performance in grade five would be internalized by a K-8 or K-6 school. Moreover, the theory predicts that the effort disparity between any two configurations should be increasing in the shared grade.

To assess the strength of such distortions empirically, one could simply compare grade five scores across different configurations, though this would be unsuitable if schools with different grade spans differed along unobserved dimensions, such as the degree to which parents invest in their children's education. A difference-in-differences approach, making this comparison across configurations both before and after the accountability reform's implementation, would control for any time-invariant differences. In my preferred approach, I adopt a triple-differences estimation strategy, comparing the difference-in-differences estimates across grades (for instance, grade five versus four) to account for differentially trending unobservables. Ratchet effects are then identified under the plausible assumption that unobservables do not differentially trend both across configurations and by grade over time.

Applying this triple-differences approach, my analysis reveals sizable distortions across K-5 and K-8 schools – between 4.7 and 5.9 percent of a standard deviation in the grade five score in favor of K-5 schools. The analogous distortion for the comparison between K-5 and K-6 schools is between 3.9 and 5.6 percent of a standard deviation in the grade five score. To place these significant effects into context, the literature suggests that teachers account for between 8 and 15 percent of a standard deviation in test scores (see Rivkin, Hanushek, and Kain 2005; Rothstein 2010; Chetty, Friedman, and Rockoff 2013a). My findings are consistent with the predictions of the model – that effort and scores will be higher at schools with shorter horizons, and that the disparity between shorter- and longer-horizon schools will be increasing in the grade.

The results are obtained without having to make overly-restrictive identifying assumptions and are robust to the most serious concerns about validity. In particular, I reject the possibility that they are driven by supply-side changes in the grade configuration of schools, differential household sorting across configurations and grades over time, location differences (rural, suburban, urban) across configurations, or the introduction of subsequent reforms.

While I take my results as reflecting differences in teacher effort, I also assess a rival dynamic gaming mechanism, whereby teachers are re-sorted across grades by the school

principal according to their teaching ability. My findings indicate that, while teacher sorting by principals is important, differential effort is likely to be the primary channel though which such gaming occurs, as it accounts for more than half of the total estimated ratchet effect.[5]

This study of ratcheting behavior is relevant to a broad class of incentive schemes that are employed in education settings and beyond, which condition on the prior decisions of agents in order to account for heterogeneity in inputs.[6] These include systems that evaluate absolute and relative performance alike, where relative performance systems feature targets which are determined by multiple agents in the system (for instance, bonus receipt for heterogeneous workers competing against each other in a tournament). My estimates demonstrate that there is a clear tradeoff when conditioning targets: while efficiency is increased as agents are held less accountable for factors beyond their control, nontrivial distortions are likely to arise when future targets can be manipulated. The theoretical conditions I derive suggest a way forward: dynamic distortions can be reduced by lowering the target at the cost of some contemporaneous efficiency, a finding that policymakers should be cognizant of when designing incentive schemes.[7]

The rest of the paper is organized as follows: The next section presents a simple theoretical model of dynamic gaming that yields the main insight used subsequently to estimate dynamic distortions. Section III describes the data, presenting stylized facts regarding the aggregate impact of the North Carolina incentive reform. Section IV outlines the empirical strategy, and Section V reports the main results, along with a set of robustness checks that address the key threats to validity. Section VI then explores likely mechanisms underlying the estimated dynamic effects, and Section VII concludes.

[5]A secondary contribution of this paper is the identification of *time-varying* classroom effects at the configuration-grade level, such as teacher effort and re-allocation, using raw score data along with features of the underlying incentive environment. Notably, in the former case, I am able to do so without relying on generally poor proxies for effort. This builds upon an established literature concerning the inference of time-invariant teacher effects, often referred to as teacher ability or quality, from such data (see Kane and Staiger 2001, 2008; Todd and Wolpin 2003, 2007; McCaffrey et al. 2004; Rothstein 2010).

[6]There is a small empirical literature measuring ratchet effects outside of education. Cooper et al. (1999) and Charness, Kuhn, and Villeval (2010) provide evidence of these effects within a simple experimental environment. Allen and Lueck (1999) and Parent (1999) present suggestive observational evidence using cross-sectional variation and, in the latter case, without information on the nature of high-powered pay or targets.

[7]Ultimately, the optimal target will depend on the long-run effects of the distortions. This is a difficult but potentially important issue to address, given the persistence of teacher effects established in Chetty et al. (2013b). A careful exploration of it remains for future investigation.

II. A Stylized Model

To develop intuition as to the possible workings of the ratchet effect[8] in a realistic setting, I extend the dynamic moral hazard literature in this section. In particular, I focus on the strand which explores ratcheting behavior when the planner commits to a suboptimal incentive scheme with a well-specified revision procedure (see Weitzman 1980; Holmstrom 1982; Keren, Miller, and Thornton 1983).[9]

My first extension is to introduce a finite-period setting, which is motivated by the dependence of school-level targets on the prior test scores of students who do not attend the same school forever. My formulation also departs from the existing literature in that output in the model depends on inputs in the current period and all prior periods according to a production function with an evolving educational capital stock (described more fully in the next subsection).[10] Given this relationship, the current choice will affect future output levels even if the target does not depend on the prior score. By modeling the production technology, the degree of dynamic gaming then depends on the extent to which the target coefficient deviates from the growth in output over time (rather than zero). Guided by the institutional details of the educational accountability system implemented throughout North Carolina in the 1996-97 school year,[11] the theory yields a new insight regarding the identification of ratchet effects, in addition to several testable predictions for the empirical investigation that follows.

[8] In general, a ratchet effect arises if the high-powered target for the next period depends on the output level in the current period. If this is the case, then any contemporaneous increase in productivity results in a one-time heightened benefit, but also permanently raises the bar for future monetary rewards, causing agents to adjust their behavior in response.

[9] A simple treatment along the lines of Weitzman (1980) is presented in Appendix A.1 (available online). See also Freixas, Guesnerie, and Tirole (1985), Lazear (1986), Baron and Besanko (1987), Gibbons (1987), Laffont and Tirole (1988), and Kanemoto and MacLeod (1992), which consider the ratchet effect under mechanisms with limited or no commitment.

[10] The period-specific 'capital' stock measures each student's ability to learn in the given period. It depends on the innate ability of the student and all of the educational inputs that she has faced prior to that point in time. This is appropriate given the cumulative nature of the education process.

[11] See Appendix B.1 (available online) for additional detail on the reform.

II.A. The Environment

Agents and Actions

Given that the incentive scheme under the accountability reform consists of grade-specific targets for each school, it is natural to focus on school principals as agents in the model. The principal is assumed to observe the test scores associated with each teacher and to be able to calculate the school-level target, a relatively straightforward exercise since the target is equal to a given coefficient α multiplied by the prior score. Using this information, she coordinates the actions of all teachers through monitoring and (potentially) sanctions to maximize the school's payoff. I abstract away from intra-school incentives in the model.[12]

Suppose there are S schools, indexed by $s \in \{1, \ldots, S\}$, and let the grade within a particular school be referenced by $g \in \mathcal{G}_c = \{0, \ldots, G_c\}$, where G_c is the last grade served by schools with grade configuration c, normalized so that $g = 1$ is the first grade with high-powered incentives attached.[13] In any given year t, each school s with a finite horizon dictated by its configuration c chooses a set of grade-specific effort levels $\{e_{scgt}\}_{g \in \mathcal{G}_c}$. Each choice e_{scgt} is an input in the production of educational achievement for students and is selected from $\mathbb{R}^+$ according to the school's payoff.

Inputs and Production Technology

For simplicity, I model a single representative subject, abstracting away from the two tested subjects used in practice in North Carolina.[14] At the end of every year t, a test is written in this subject by all students in school s, generating average test scores for each school-grade pair. These scores are denoted by y_{scgt} and are taken to be a measure of educational output for the relevant group of students and the representative teacher for that grade.

[12]This modeling choice is made to focus on the core idea of ratcheting behavior. It assumes that the principal is capable of perfectly coordinating her teachers. As coordination becomes more imperfect, ratchet effects would naturally be weakened.

[13]For example, according to this notation, given that the receipt of the bonus in North Carolina depends on the scores for grades three through eight, $g = 0$ corresponds to grade two and $g = G_c = 3$ corresponds to grade five for a K-5 school (while I do not focus on earlier low-stakes grades here, grade one would be represented by $g = -1$ in this case). For a 6-8 school, there is no grade for which $g = 0$, as it represents grade five at a different school. Thus, $g \in \mathcal{G}_c = \{1, 2, 3\}$ for such middle schools.

[14]The representative subject assumption can be made without loss of generality, since the dependence of bonus receipt on composite measures implies that dynamic effects will be manifested in both scores.

Education is inherently cumulative, with learning in each period building upon what came before. I capture this using the concept of 'educational capital,' defining it to be the stock of skills and knowledge a student has accumulated up to a given time for the purpose of learning. It reflects the idea that inputs to learning, such as the student's raw intelligence and the contributions of her teachers, have a lasting impact on her capacity to learn in the future. As these prior inputs are not directly observed, I summarize the prior end-of-grade educational capital which students begin grade g with using the prior score, $y_{scg-1t-1}$.

Given this definition, I model the score y_{scgt} as depending on the effort e_{scgt} exerted by the representative teacher for the school-grade pair, the ability of the teacher a_{scgt}, the prior end-of-grade educational capital for current grade g students $y_{scg-1t-1}$, and a grade-school-year shock u_{scgt}. Teacher effort and shocks are treated as common to all students within a classroom – a reasonable assumption given that the average outcome for each grade is what matters for satisfying the school-level target. In addition, teacher effort is modeled exclusively as the representative teacher's contribution to the average score of her students, meaning that I abstract away from multiple tasks, such as devoting effort to disciplining students. I also consider the effect of teacher effort to be permanent so that it affects the subsequent score in the same way as educational capital. In general, let the student's score in school s, grade configuration c, grade g and time t be given by

$$y_{scgt} = H(y_{scg-1t-1}, e_{scgt}, a_{scgt}) + u_{scgt}\,,$$

which potentially allows for teacher effort and the capital stock of the average student to interact in the production of learning. To develop intuition and make the identification strategy that follows more transparent, I focus on a linear specification – a standard assumption made in the educational literature. Under the linear technology, the score is given by

$$y_{scgt} = \gamma y_{scg-1t-1} + e_{scgt} + a_{scgt} + u_{scgt}\,. \tag{1}$$

Incentives and Preferences

Suppose, as is the case for the North Carolina reform, that the planner selects an incentive scheme that rewards teachers at a school with a monetary bonus b if the school-level score exceeds the target. Given the average scores y_{scgt} and targets $\hat{y}_{scgt} \equiv \alpha y_{scg-1t-1}$ for each grade

within the school,[15] this award criterion is equivalent to the sum of the scores exceeding the sum of the targets across grades.

The choice of effort for each grade g and time t depends on the positive inter-temporal depreciation rate δ, the probability of receiving the monetary bonus b and the convex cost $C(\cdot)$ of the effort that is exerted. Therefore, the payoff function for an infinitely-lived school s serving G_c grades at time t is

$$U_{sct} = \sum_{t=1}^{\infty} \delta^{t-1} \left\{ b \left[1 - F_c \left(\sum_{g=1}^{G_c} ((\alpha - \gamma) y_{scg-1t-1} - e_{scgt} - a_{scgt}) \right) \right] - \sum_{g=0}^{G_c} C(e_{scgt}) \right\} \tag{2}$$

where $F_c(\cdot)$ is the cdf of the sum of grade-specific shocks ($\sum_{g=1}^{G_c} u_{scgt} \equiv G_c \bar{u}_{sct}$), and the benefit portion of the payoff function arises from the probability of receiving the bonus $Pr[\sum_{g=1}^{G_c} y_{scgt} > \sum_{g=1}^{G_c} \hat{y}_{scgt}]$, which (using equation (1)) is equivalent to $Pr[\sum_{g=1}^{G_c} u_{scgt} > \sum_{g=1}^{G_c} ((\alpha - \gamma) y_{scg-1t-1} - e_{scgt} - a_{scgt})]$.

II.B. Optimal Effort Levels

Given the technology in equation (1), the problem for school s at time t is to choose the stream of effort levels $\{\{e_{scgt}\}_{g \in \mathcal{G}_c}\}_{t=1}^{\infty}$ to maximize the objective in equation (2). Defining $\Pi_{sct} \equiv -\sum_{g=1}^{G_c} (e_{scgt} + a_{scgt} + (\gamma - \alpha) y_{scg-1t-1})$ and assuming a quadratic cost function $C(e) = \frac{d}{2} e^2$, the first-order conditions that govern these choices are given by

$$\frac{d}{b} e_{scgt} = \begin{cases} f_c(\Pi_{sct}) + \delta(\gamma - \alpha) \sum_{i=0}^{G_c - g - 1} \delta^i \gamma^i f_c(\Pi_{sct+1+i}) & \text{for } 1 \leq g < G_c \\ f_c(\Pi_{sct}) & \text{for } g = G_c \end{cases},$$

where the second term on the right hand side of the equation for $1 \leq g < G_c$ is the distortion due to dynamic gaming, $f_c(\cdot)$ is the pdf of $\sum_{g=1}^{G_c} u_{scgt}$ and $f_c(\Pi_{sct})$ represents the school-specific contemporaneous incentives for period t unrelated to the ratchet effect.[16] While these conditions cannot be used to solve for each optimal effort level explicitly, they do allow the relationship between effort levels in consecutive grades to be characterized.

For the remainder of this subsection, I assume that $\delta > 0$, $\gamma > 0$ and the high-powered

[15] Note that the multiplicative coefficient α is positive in my empirical application, as it is derived by regressing a current positive test score on a smaller positive prior one.

[16] Such incentives affect the degree to which teacher effort matters at the margin for receiving a bonus. They are generated by contemporaneous target imperfections, such as the scheme failing to account for transitory shocks or grade-to-grade differences in teacher ability.

target coefficient exceeds the growth rate of the score ($\alpha > \gamma$).[17]

Lemma 1 *Effort is weakly increasing in the grade g.*

The proof is contained in Appendix C.1 (available online). As the effort choice affects a larger number of future targets and the targets grow at a faster rate than the score ($\alpha > \gamma$), teachers are increasingly penalized for exerting higher effort. Thus it is optimal to select a lower level of effort as the horizon increases (g is further away from the final grade offered G_c). For similar reasons, the converse is also true: effort is weakly decreasing in g if target growth is outpaced by score growth ($\alpha < \gamma$).

To compare grade g outcomes for two different grade structure types, closed-form solutions for effort cannot be derived from the general first-order conditions. Thus, I provide intuition for the empirical analysis that follows by making an additional simplifying assumption that the incentive scheme is linear, in which case, the nonlinear Π terms drop away, leaving only ratchet effects that differ according to the school configuration and leading to expressions that are analytically tractable.[18] The conditions become

$$e_{cg} = \begin{cases} \frac{b}{d}\left[1 + \delta(\gamma - \alpha)\sum_{i=0}^{G_c - g - 1} \delta^i \gamma^i\right] & \text{for } 1 \leq g < G_c \\ \frac{b}{d} & \text{for } g = G_c \end{cases}.$$

Proposition 1 *Assuming that initial educational capital stock and teacher ability are identical across two school configurations c and c', such that one school serves a greater number of grades ($G_{c'} > G_c$), the test score for any particular grade g will be greater at the school serving fewer grades ($y_{cg} > y_{c'g}, \forall\, g \in \mathcal{G}_c$).*

The proof is contained in Appendix C.2 (available online). To interpret Proposition 1, consider the following example of a pair of typical K-5 and K-8 schools in North Carolina. Using the notation of the model, the K-5 and K-8 schools serve $G_c = 3$ and $G_{c'} = 6$ grades with high-powered incentives attached, respectively. As shown in the proof, the first-order

[17] The predictions that follow would be reversed if target growth is outpaced by score growth ($\gamma > \alpha$). Based on the empirical results in this paper and related structural work, I do not believe this to be the case.

[18] Such an assumption is made for expositional convenience and is not necessary for the propositions that follow. They continue to hold under a nonlinear scheme if relatively mild assumptions are imposed concerning the correlation of shocks and similarity of target imperfections across grades (details available upon request).

conditions imply that the dynamic distortion for a particular grade is always smaller for the school with the shorter horizon, which is the K-5 school in this case. Intuitively, K-8 schools always have a greater number of future grades to consider when determining their effort decision in grades three, four or five. Figure A.1 (available online) illustrates this comparison, where the effort level at the K-5 school is higher than at the K-8 school for each grade shared by the two configurations. Combined with the assumptions stated in Proposition 1, this pattern in effort also holds for test scores, as illustrated in Figure A.2 (available online). An analogous result holds for a comparison between K-5 and K-6 schools.

Proposition 2 *Under the stated assumptions of Proposition 1 and assuming* $\delta\gamma < 1$*, the positive difference between* y_{cg} *and* $y_{c'g}$ *is increasing in* g*,* $\forall\ g \in \mathcal{G}_c$*.*

The proof is contained in Appendix C.3 (available online). Using the same comparison of K-5 and K-8 schools, Proposition 2 implies that distortions diminish at a faster rate for K-5 schools when moving from one grade to the next higher grade. Combining Propositions 1 and 2, the score differential between K-5 and K-8 schools is predicted to be positive in favor of the former type for each shared grade, and this difference should be greatest for grade five – this result is reflected in Figure A.2 (available online) and is the main hypothesis to be tested empirically.[19] Given the preceding theoretical predictions, I now turn to the data used in my empirical analysis.

III. Data and Descriptive Statistics

To determine whether conditioning targets on prior scores leads to distortions of effort across grades, I utilize a rich longitudinal data set provided by the North Carolina Education Research Data Center (NCERDC). This includes information on North Carolina students, teachers and schools for the years 1994 through 2005.[20,21] Given that the accountability

[19] Given the general nonlinear first-order conditions at the beginning of this subsection, it is clear that the dynamic gaming effects are attenuated by lowering the target coefficient α toward the growth rate γ. Indeed, they are eliminated altogether if the planner sets $\alpha = \gamma$. Intuitively, such a target coefficient no longer punishes teachers in future for exerting higher effort today. However, there is a tradeoff associated with implementing this prescription for eliminating the distortions, as contemporaneous incentives are potentially weakened when the target becomes easier to satisfy.

[20] See Appendix D.1 (available online) for greater detail on the available data.

[21] For expositional convenience, I refer to academic years using the calendar year in which they end. For instance, 1994 refers to the 1993-94 school year.

reform took effect in 1997, I refer to 1994, 1995 and 1996 as pre-reform years, and 1997 through 2005 as post-reform years. The data set contains yearly standardized test scores for each student in mathematics and reading from grades two to eight.[22] These scores are comparable across time and grades through the use of a developmental scale.[23] Using this scale and unique encrypted identifiers, the progress of individual students can be tracked over their educational careers.[24] The data set also links students to their teacher and school in each year for grades three through eight.

In addition to student scores, the data provide extensive student, teacher and school characteristics. For the purposes of this study, the most important student observables are parental education, ethnicity, and exceptionality classifications. With regard to teachers, the relevant characteristics are the score on the test used to obtain a teaching license and the number of years of teaching experience. The data set also contains information on the location for each school, using five classifications ranging from a large city to a rural area, the proportion of students eligible for a free or reduced-price lunch, the number of years that the principal has been in charge of a given school, the number of classes by grade offered by a school, and – especially relevant for this study – each school's grade configuration.

With respect to the distribution of schools by grade structure, there are 849 K-5 schools, 97 K-8 schools, 102 6-8 schools and 104 K-6 schools in the sample. These tallies are approximate, as a subset of schools open, close or switch configuration during the period of study. The K-5, K-8 and K-6 counts are 661, 78 and 36 respectively for those that do not switch at any point during the period of interest. The strong decline in K-6 schools comparing the less and more restrictive samples can be attributed to the fact that many of those open in the pre-reform period switched to a K-5 configuration early in the post-reform period. Given the relatively small number of K-6 schools that do not switch, one would expect diminished statistical power when analyzing gaming behavior using K-5 and K-6 schools. This should be kept in mind when interpreting the results.

[22]What are referred to as 'grade two' tests are administered in September of the grade three year. All other tests are administered in May or June of the school year.

[23]Each point on the developmental scale is designed to measure the same amount of learning, regardless of the grade to which the score corresponds.

[24]This clear interpretation of learning over time is the motivation for basing my analysis on such scale scores. However, all results are robust to using scores that are standardized at the grade-year level instead.

Table 1
Descriptive Statistics

	All Configurations (pre and post)		By Configuration (pre only)		
Variable:	Mean	St. Dev.	K5	K6	K8
Combined Test Score					
Grade 3	291.5	18.8	286.3	283.7	285.7
Grade 4	303.3	18.3	297.6	294.8	296.5
Grade 5	314.7	16.9	308.0	305.6	307.0
Student - Parental Education					
No High School	0.10	0.30	0.11	0.13	0.14
High School Graduate	0.42	0.49	0.42	0.46	0.47
Trade School	0.09	0.28	0.05	0.05	0.04
Community College	0.11	0.32	0.14	0.15	0.16
4-Year College	0.22	0.42	0.22	0.18	0.14
Graduate Degree	0.06	0.23	0.06	0.04	0.04
Student - Ethnicity					
White	0.64	0.48	0.68	0.66	0.82
Black	0.28	0.45	0.28	0.28	0.13
Other	0.08	0.27	0.04	0.05	0.05
Student - Exceptionality					
Learning Impairment	0.12	0.32	0.12	0.11	0.10
No Special Label	0.74	0.44	0.76	0.78	0.79
Gifted	0.14	0.35	0.12	0.11	0.11
School					
Prop. Free Lunch Eligible	0.38	0.20	0.36	0.38	0.34
Principal Tenure in School†	3.3	2.2	1.4	1.4	1.4
Classes Per Grade (Gr. 3-5)†	3.5	1.4	3.4	3.0	1.9
Teacher Experience†	12.9	6.0	13.5	12.8	13.2
Teacher Licensure Test Score†	0.00	0.55	0.00	-0.05	-0.01
Locale - Large City	0.06	0.23	0.02	0.09	0.00
Locale - Mid-Size City	0.21	0.41	0.32	0.22	0.01
Locale - Large Suburban	0.05	0.21	0.03	0.06	0.00
Locale - Mid-Size Suburban	0.13	0.34	0.17	0.09	0.05
Locale - Small Town & Rural	0.55	0.50	0.46	0.55	0.94

NOTE.—Student statistics are averaged across all students from 1994 to 2005, while teacher and school statistics are averaged at the school level over the same period († indicates no data for 1994). School location and student categories are both mutually exclusive and exhaustive.

Descriptive statistics for the remaining variables of interest are provided in Table 1. Student scores and characteristics are presented at the student level from 1994 to 2005, while teacher and school statistics are averaged at the school level over the same period. I also report the mean of each variable in the pre-reform period by school configuration. The statistics show that K-8 and K-6 schools are observably different from their K-5 counterparts along dimensions such as test scores, parental education and race. These disparities are mainly

due to the fact that K-8 and K-6 schools are disproportionately located in rural areas, while K-5 schools tend to be found in rural as well as urban and suburban areas.[25] Accounting for the school's locale in the empirical analysis is therefore likely to be important.

Beyond basic statistics, Appendix E (available online) presents graphical evidence to show that test scores increase from the pre- to post-reform period for all tested grades. This pattern is in keeping with a positive overall effect of the reform on student achievement. Moreover, the growth in scores over time is monotonically increasing in the grade, which is the type of dynamic pattern predicted by the theoretical model. Such growth is also not uniform across school configurations, with K-5 schools realizing the largest gains in grade five (as predicted by Proposition 1). With this suggestive evidence in hand, I now set out my basic econometric strategy to test for ratchet effects in a formal way.

IV. Empirical Strategy

The theoretical analysis draws attention to a method for identifying ratchet effects using variation in the horizon a school faces. In particular, Proposition 1, which states that the average score will be higher in a given grade at a school serving fewer grades, is testable under the assumption that schools with different grade configurations are otherwise identical. For several reasons, the condition that grade spans are exogenous is unlikely to be satisfied in practice. I briefly discuss why this is the case, before detailing my strategy for dealing with unobserved differences across schools.

Owing to a variety of historical factors, the popularity of different elementary school grade configurations has waxed and waned over time, potentially leading such configurations to be non-randomly represented in the current population of schools.[26] As a result, there is ample

[25]For the full sample, approximately 396 K-5, 87 K-8 and 71 K-6 schools are located in rural areas. For the subsample of schools that do not switch grade configuration, the counts are 297, 69 and 30.

[26]In the early twentieth century, K-8 schools were the dominant structure in the United States. In an effort to ease the transition between elementary and secondary school and alleviate enrollment pressures arising from immigration flows, K-6 and junior high schools became more prevalent as the century progressed. In the 1960s, research indicating that students were maturing earlier caused policymakers to shift grade six from K-6 schools to the junior high structure, leading to the creation of K-5 and 6-8 configurations. However, middle schools began to fall out of favor in the 1980s and 1990s as the large institutions were perceived to be inadequately serving their students. Later research, including survey evidence by Juvonen et al. (2004) and empirical analyses by Alspaugh (1998), Hanushek, Kain, and Rivkin (2004) and Rockoff and Lockwood (2010), also suggested that a higher number of school transitions was deleterious to student development.

reason to believe that a disparity in scores between two schools with different horizons reflects more than just differential ratchet effects. For instance, the distribution of student ability may differ across K-5, K-6 and K-8 schools. If this is the case, then each configuration may be associated with a different initial level of educational capital in the production process, leading to disparities in subsequent scores regardless of whether incentives vary according to the school's horizon. Similarly, if the quality of teachers, surrounding neighborhood characteristics or educational resources differ by school type, variation in scores across grade configurations may be incorrectly interpreted as evidence of dynamic gaming.

To isolate the variation in scores that may arise from dynamic incentives, I begin by considering a difference-in-differences approach, using pre-reform scores as a baseline to control for unobserved factors that vary across different grade spans. In order to compare the grade five score between K-5 and K-8 schools, for example, I would simply construct the difference-in-differences score

$$\Delta\Delta y_{\text{K5}-\text{K8},5,post-pre} = (y_{\text{K5},5,post} - y_{\text{K5},5,pre}) - (y_{\text{K8},5,post} - y_{\text{K8},5,pre}) \,.$$

Such an approach adjusts for both pre-existing disparities and shared changes (common trends) between school configurations in inputs and the production process. If incentives are the only time-varying factor leading to differential changes over time and the underlying technology is linear, then the technique will produce an unbiased estimate of the dynamic gaming distortion.

Although the former assumption is significantly less restrictive than simply controlling for observable characteristics, the strategy remains susceptible to differentially trending variables which are unrelated to incentives. For example, if families sort across neighborhoods or teachers sort across schools, then the composition of educational production inputs might evolve over time. My initial strategy accounts for this possibility by conditioning on observed student, teacher and school controls X_{sgt} prior to computing difference-in-differences estimates. As there are many such estimates to consider, I first estimate the equation

$$y_{scgt} = X'_{sgt}\beta + \sum_{c=1}^{C} \sum_{g \in \mathcal{G}_c} (\phi_{c,g,pre} + \phi_{c,g,post}) + \varepsilon_{scgt} \tag{3}$$

where each ϕ is an interacted indicator variable that adjusts the score for every combination

of grade, school type and period.[27] In essence, each fixed effect is a score for a particular school configuration and grade in the pre- or post-reform period, adjusted for the vector of observable controls.

Upon estimating equation (3), I use F-tests of the relevant ϕ coefficients to recover difference-in-differences estimates of the adjusted score for each grade. For instance, the estimate comparing grade g scores between K-5 and K-8 schools is

$$\Phi_{\text{K5}-\text{K8},g,post-pre} \equiv (\phi_{\text{K5},g,post} - \phi_{\text{K5},g,pre}) - (\phi_{\text{K8},g,post} - \phi_{\text{K8},g,pre}) \,. \tag{4}$$

If unobserved trends are common across grade configurations, then $\Phi_{\text{K5}-\text{K8},g,post-pre} > 0$ satisfies the criterion for dynamic gaming behavior as in Proposition 1.

Despite the merits of the proposed difference-in-differences strategy, differentially trending unobservables may bias the estimates. Potential areas of concern include demand-side sorting by households or teachers across schools of different grade configuration[28] and supply-side changes in the distribution of school configurations over time.[29] One approach for addressing the supply-side issue is to restrict the difference-in-differences analysis to the subset of schools that maintain the same grade configuration during the period of interest, which I do in the following section. However, assuming schools compete with each other locally, selection bias may persist due to the competitive effects of schools that switch on non-switching ones.[30]

The most robust way to address the preceding identification issues is to employ a triple-differences approach. Given that difference-in-differences estimates can be computed for every grade that is shared by any two school configurations, a triple difference can be formed using the difference between such estimates for any two grades. For instance, the estimate

[27]The results do not appreciably change when allowing for control coefficients to vary by grade (β_g) or including school-level fixed effects.

[28]Since K-6 and K-8 schools are predominantly found in rural areas, upward bias would result if shifting economic conditions cause low-ability households to differentially sort into rural areas. Differential changes in unobserved district salary schedules might also lead to bias from teacher sorting.

[29]As discussed in Section III, North Carolina policymakers increasingly shifted toward the K-5/6-8 model during the post-reform period. If the schools were systematically selected for this transition based on unobserved determinants of performance, bias might result. Given that the data indicates switching K-8 schools have a lower score than those that do not switch, the direction of such bias is likely downward.

[30]To see why, consider a district with two K-8 schools, one of which is underperforming, and the other, high-performing. If the underperforming one converts to a K-5 school and such a configuration is more desirable than a K-8 one, then the new school may attract some higher ability students from the previously high-performing K-8 school, resulting in upward-biased estimates.

comparing grade four and five scores between K-5 and K-8 schools is

$$\Phi_{\text{K5-K8,5-4},post-pre} = \Phi_{\text{K5-K8,5},post-pre} - \Phi_{\text{K5-K8,4},post-pre}\,, \tag{5}$$

where the difference-in-differences estimates $\Phi_{\text{K5-K8,5},post-pre}$ and $\Phi_{\text{K5-K8,4},post-pre}$ are defined by equation (4).

Such an analysis controls for time-invariant effects and shared trends between configurations, but also accounts for differentially trending unobservables as long as their effect is grade-invariant. If one believes that household and teacher sorting and evolving school competition do not affect scores differentially by configuration and grade, then remaining demand- or supply-side selection bias is addressed by the triple-differences approach. A finding of $\Phi_{\text{K5-K8,5-4},post-pre} > 0$ is interpreted as satisfying the criterion for dynamic gaming behavior as in Proposition 2, which predicts that the magnitude of dynamic distortions is increasing in the grade. I now turn to the difference-in-differences and triple-differences estimates to determine whether the data are consistent with ratcheting behavior.

V. Results

Figures A.3 and A.5 (both available online) already provide preliminary evidence consistent with dynamic gaming. I now analyze these effects in a more formal way empirically. In particular, I estimate equation (3) under three different specifications, depending on the components of the control vector X_{sgt}. These specifications are defined in Table A.1 (available online), where the coefficients of each regressor are reported. Specification (1) uses the raw score without controls, while specification (2) includes student characteristics (such as the ethnicity of students, the education of their parents and their exceptionality classification), the school-level proportion of students who are eligible for the free lunch program and controls for the locale of the school. Specification (3) then adds the licensure test score of each student's teacher.

All coefficients are significant and of the expected sign. A higher combined test score in mathematics and reading is associated with students who are white, who have parents with a more advanced education and who are labeled as being exceptional. The score is also positively related to students attending a school with a lower free lunch participation rate

and those with teachers who scored higher on their licensing test.

For specifications (1) through (3) in Table A.1 and grades three through five, I transform the relevant fixed effects from equation (3) into first-difference, difference-in-differences and triple-differences estimates, as in equations (4) and (5). The results for K-5 and K-8 schools, and K-5 and K-6 schools are reported in Table 2. In every case, the difference between pre- and post-reform scores for a specific configuration is positive and significant, consistent with the descriptive evidence. Using specification (3), the pre-to-post gain in grade five scores for K-5, K-8 and K-6 schools is 8.8, 7.3 and 5.9 developmental scale points, respectively. The analogous gains in grade four scores are 7.5, 7.1 and 5.6 points and in grade three scores are 6.8, 6.7 and 4.9 points.[31] This highlights the fact that score growth increases with the grade regardless of the school's grade configuration.[32]

The more interesting results with regard to ratchet effects are the difference-in-differences and triple-differences estimates. For the comparison between K-5 and K-8 schools, the difference-in-differences estimates reported in Table 2 are statistically indistinguishable from zero for each grade when no observable controls are included. However, after introducing controls, the grade five estimates are positive and significant, which is consistent with Proposition 1.[33] That is, controlling for trending observables and the pre-reform outcome, the school with the shorter grade horizon (K-5) has a higher score. Moving on to the preferred triple-differences strategy, the corresponding estimates are positive and significant across all three specifications when comparing grade five to four and positive but smaller for the comparison between grades four and three when including controls. This pattern across grades is precisely what is predicted by the theory (Proposition 2).

The magnitude of dynamic distortions suggested by the difference-in-differences and triple-differences estimates is substantial. Comparing K-5 and K-8 schools, the differential effect of the scheme is estimated to be between 1.46 and 1.53 developmental scale points for grade five, depending on the control-based specification used. This is equivalent to an effect that

[31] Counterintuitively, gains are always lowest for K-6 schools. This reflects selection bias arising from K-6 schools disproportionately switching to a different configuration. This issue is addressed in Table 3.

[32] Given that the standard deviation of the score is larger in grade three and four than the 16.9 developmental points in grade five (see Table 1), this pattern becomes even more pronounced when adjusting for variation in scores.

[33] The estimates for grades three and four are also positive, but not significantly so.

Table 2
Main Results

	c = K5 vs. c' = K8			c = K5 vs. c' = K6		
Specification:	(1)	(2)	(3)	(1)	(2)	(3)
Grade 5 DinD						
$\Phi_{c,5,post-pre}$	9.01***	9.51***	8.77***	9.01***	9.51***	8.77***
	(0.22)	(0.15)	(0.14)	(0.22)	(0.15)	(0.14)
$\Phi_{c',5,post-pre}$	8.41***	7.98***	7.31***	7.30***	6.76***	5.92***
	(0.36)	(0.29)	(0.31)	(0.45)	(0.29)	(0.32)
$\Phi_{c-c',5,post-pre}$	0.60	1.53***	1.46***	1.71***	2.75***	2.84***
	(0.43)	(0.33)	(0.34)	(0.52)	(0.33)	(0.35)
Grade 4 DinD						
$\Phi_{c,4,post-pre}$	7.76***	8.10***	7.52***	7.76***	8.10***	7.52***
	(0.19)	(0.13)	(0.14)	(0.19)	(0.13)	(0.14)
$\Phi_{c',4,post-pre}$	7.96***	7.50***	7.05***	6.29***	6.02***	5.62***
	(0.43)	(0.36)	(0.42)	(0.51)	(0.33)	(0.37)
$\Phi_{c-c',4,post-pre}$	-0.20	0.60	0.47	1.47***	2.08***	1.89***
	(0.47)	(0.38)	(0.45)	(0.56)	(0.36)	(0.40)
Grade 3 DinD						
$\Phi_{c,3,post-pre}$	7.21***	7.48***	6.79***	7.21***	7.48***	6.79***
	(0.20)	(0.15)	(0.16)	(0.20)	(0.15)	(0.16)
$\Phi_{c',3,post-pre}$	7.26***	7.27***	6.74***	5.96***	5.92***	4.94***
	(0.46)	(0.40)	(0.41)	(0.48)	(0.33)	(0.38)
$\Phi_{c-c',3,post-pre}$	-0.05	0.21	0.04	1.25**	1.55***	1.85***
	(0.51)	(0.43)	(0.45)	(0.53)	(0.36)	(0.41)
Triple Differences						
$\Phi_{c-c',5-4,post-pre}$	0.80**	0.93***	0.99**	0.24	0.66**	0.95***
	(0.38)	(0.34)	(0.44)	(0.31)	(0.27)	(0.36)
$\Phi_{c-c',4-3,post-pre}$	-0.15	0.39	0.42	0.22	0.53	0.04
	(0.38)	(0.41)	(0.47)	(0.33)	(0.34)	(0.44)

NOTE.—For each specification defined in Table A.1 (and in words in the first paragraph of Section V), this table reports first-differences, difference-in-differences and triple-differences estimates constructed from joint F-tests of the interaction dummies included in the regression. Standard errors adjusted for clustering at the school level are reported in parentheses.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

is between 8.6 and 9.1 percent of a standard deviation in the grade five score. For the triple differences estimates, the effect is estimated to be between 0.80 and 0.99 scale points for the grade five to four comparison (or between 4.7 and 5.9 percent of a standard deviation in the grade five score). The analogous difference-in-differences estimate for the comparison between K-5 and K-6 schools is between 1.71 and 2.84 scale points, while the analogous triple-differences estimate is between 0.66 and 0.95 scale points (or between 3.9 and 5.6

percent of a standard deviation in the grade five score).[34]

V.A. Robustness

The astute reader may note that I have yet to highlight the comparison between K-6 and K-8 schools. Rather than an oversight, this decision stems from a lack of power, owing to relatively few observations for both grade configurations in the sample. Propositions 1 and 2 intuitively predict that the difference-in-differences and triple-differences estimates for this comparison should be smaller than analogous results for the K-5 and K-8 comparison and larger than those for the K-5 and K-6 comparison. While not directly reported, the estimates for K-6 and K-8 schools can be computed from Table 2 by taking the difference between the estimates for the other two comparisons. However, doing so fails to account for a previously mentioned issue that is particularly important in the case of K-6 and K-8 schools and compounded when comparing the two: differential supply-side changes in the distribution of schools by configuration and grade.

To address this supply-side validity concern for all comparisons, Table 3 reports difference-in-differences and triple-differences results for the full sample of schools (as in Table 2) and for the subsample of schools that maintain their grade configuration throughout the pre- and post-reform periods. Under the subsample restriction, the grade five difference-in-differences estimate with controls diminishes only slightly for the comparison between K-5 and K-8 schools and more so for the K-5 and K-6 comparison. This makes the former and latter estimates statistically indistinguishable from each other. However, each estimate is still separately significant. The grade five to four triple-differences estimates change more substantially across specifications, with those for the K-5 and K-8 comparison increasing in significance and rising to between 1.43 and 1.66 scale points, and those for the K-5 and K-6 comparison becoming insignificant but remaining positive for schools with a stable configuration. Using these results, analogous estimates for the K-6 and K-8 comparison are between 0.51 and 0.83 scale points. These point estimates are lower than the results for the

[34]Placing these results in context, a child of a college-educated parent is predicted to score 45.1 percent of a standard deviation in the grade five score higher than one whose parent does not have a college degree. In addition, the score increase that would occur by lowering poverty in a school (as measured by free lunch participation) by one standard deviation would be 6.2 percent of a standard deviation in the grade five score.

K-5 and K-8 comparison and, owing to more sizable standard errors, the hypothesis that they are larger than the results for the K-5 and K-6 comparison (as predicted by the theory) cannot be rejected. Thus, the sign, magnitude and significance of the difference-in-differences and triple-differences estimates are consistent with the dynamic gaming hypothesis.[35]

Having demonstrated the robustness of the results to supply-side changes, I now discuss and reject the most potent remaining rival hypotheses. A second potential concern is that the estimated effects are due to differential household sorting across configurations and grades over time. Such a claim is refuted for observables using Table 3 by comparing the triple-differences estimates under specification (1) and (3). In all cases, the estimates are not statistically different from each other. Given that specification (1) includes no controls while (3) includes all controls, this suggests that the results cannot be explained by sorting based on observed household characteristics. To alleviate concerns about selection on unobservables, I also compute the difference-in-differences and triple-differences of relevant observable characteristics directly. Assuming that observables and unobservables follow a similar pattern, the results in Table A.2 (available online) soundly reject the rival household sorting hypothesis. In particular, the triple-differences estimates for all parental education indicators and the free lunch poverty proportion are statistically indistinguishable from zero, suggesting the existence of common trends on a triple-differences basis.[36]

A third validity issue pertains to the differential location of schools by configuration. As noted previously, K-8 and K-6 schools are disproportionately located in rural areas of North Carolina, raising the prospect that differences in location-related characteristics may be driving the results (for instance, educators in rural areas may respond to incentives differently than their urban or suburban counterparts). Given the strength of the findings for the

[35]Lending further credence to the main dynamic gaming interpretation, I explore how the estimates of the ratchet effect evolve over time in Appendix F.1 (available online) and show that the dynamic gaming effects are most pronounced for mathematics test scores in Table A.6 (also available online), the latter of which is in keeping with the findings of multiple prior studies showing teachers have a greater effect on mathematics than on reading scores (see, for instance, Rivkin et al. 2005).

[36]Even in the absence of household sorting across grade configurations according to student ability, the movement of students across schools (whether within or across configurations) may attenuate the ratchet effect as the horizon faced by the school for that subset of students is shortened. Limited variation in student transfers provides insufficient power with which to test whether such attenuation occurs. However, given that only approximately 10 percent of students move to a different school each year, any resulting attenuation is likely to be minor and, in any case, the main results would be strengthened without it. An analogous argument applies to the approximately 12 percent of teachers who change schools each year.

Table 3
Restricted-Sample Robustness Check

	c = K5 vs. c' = K8		c = K5 vs. c' = K6	
Specification:	(1)	(3)	(1)	(3)
All Schools in Sample				
$\Phi_{c-c',5,post-pre}$	0.60	1.46***	1.71***	2.84***
	(0.43)	(0.34)	(0.52)	(0.35)
$\Phi_{c-c',4,post-pre}$	-0.20	0.47	1.47***	1.89***
	(0.47)	(0.45)	(0.56)	(0.40)
$\Phi_{c-c',3,post-pre}$	-0.05	0.04	1.25**	1.85***
	(0.51)	(0.45)	(0.53)	(0.41)
$\Phi_{c-c',5-4,post-pre}$	0.80**	0.99**	0.24	0.95***
	(0.38)	(0.44)	(0.31)	(0.36)
$\Phi_{c-c',4-3,post-pre}$	-0.15	0.42	0.22	0.04
	(0.38)	(0.47)	(0.33)	(0.44)
Stable Config Only				
$\Phi_{c-c',5,post-pre}$	0.76*	1.36***	-0.18	1.51***
	(0.44)	(0.38)	(0.64)	(0.56)
$\Phi_{c-c',4,post-pre}$	-0.67	-0.30	-1.10	0.69
	(0.47)	(0.48)	(0.77)	(0.69)
$\Phi_{c-c',3,post-pre}$	-0.78	-0.51	-0.72	0.86
	(0.53)	(0.53)	(0.84)	(0.91)
$\Phi_{c-c',5-4,post-pre}$	1.43***	1.66***	0.92	0.83
	(0.40)	(0.51)	(0.60)	(0.59)
$\Phi_{c-c',4-3,post-pre}$	0.11	0.21	-0.39	-0.18
	(0.43)	(0.54)	(0.65)	(0.92)

NOTE.—For the specification without any and with full controls, this table reports robustness checks for the difference-in-differences and triple-differences estimates by comparing the full sample results with all schools (top panel) to those for the subsample of schools that maintain a stable grade configuration over the period of interest (bottom panel). As before, the estimates are constructed from joint F-tests of the interaction dummies included in the regression. Standard errors adjusted for clustering at the school level are reported in parentheses.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

comparison between K-5 and K-8 (relative to K-5 and K-6) schools in Table 3, I report estimates for K-5 and K-8 schools by locale in Table A.3 (available online). The triple-differences estimates are positive and significant across all schools and for rural schools. However, there is insufficient power to compute urban or suburban counterparts, owing to a dearth of K-8 schools in those areas. Therefore, I construct alternative difference-in-differences estimates for K-5 schools only, which exploit the pre/post and grade (versus pre/post and configuration) dimensions, to assess the extent of dynamic gaming within each

locale. The results show that such gaming is present in rural, urban and suburban areas alike, with estimates that cannot be statistically distinguished from each other. Thus, my findings are robust to location considerations.

A fourth potential threat to validity relates to the implementation of subsequent educational reforms in North Carolina during the period of analysis. These include the introduction of charter schools to compete with conventional public schools in 1998,[37] student accountability in 2001,[38] and the federal No Child Left Behind Act in 2003. To rule out upward bias arising from these reforms differentially affecting school configurations and grades, Table A.4 (available online) presents the results of a falsification exercise where I counterfactually assume that the accountability reform began in a year other than 1997. Notably, the grade five difference-in-differences and grade five to four triple-differences estimates are largest in the actual year of the reform for the K-5 and K-8, and K-5 and K-6 comparisons.[39] Therefore, the dynamic effects that I have uncovered are robust to the implementation of additional policies during the period of interest. Having established identification, I now consider the mechanisms behind the ratchet effects.[40]

VI. Mechanisms

To clearly motivate the identification strategy, the stylized model has focused on the monitoring and coordination of teacher effort by principals as the exclusive channel for dynamic gaming to occur, setting aside alternative mechanisms for the sake of simplicity. Yet a perfectly plausible and leading rival hypothesis is that principals re-allocate teachers across grades to maximize their school's payoff, altering teacher ability (rather than effort) across classrooms. With a slight modification to accommodate the discrete nature of such a decision, the implications for within-school teacher sorting are analogous to those for effort: school principals are predicted to shift teachers with greater teaching ability to higher grades. Un-

[37] From Bifulco and Ladd (2006), the number of charter schools in 1998 was 27, growing to 67 by 2002.

[38] Fifth grade (and third grade, beginning in 2002) students were required to satisfy a specified performance threshold to advance to the next grade. Fruehwirth (2013) provides additional detail on this reform.

[39] The counterfactual point estimates in 1998 are smaller and the estimates in 2001 and 2003 are substantially and significantly smaller than in 1997. The negative estimates for 2002 onward reflect the dynamic effect in reverse, as the strong post-period effects are counterfactually attributed to the pre-period instead.

[40] For the interested reader, I also briefly discuss two ways in which the strength of these effects might be affected in Appendix F.2 (available online).

derstanding the extent to which the estimated dynamic gaming effects are due to differences in teacher effort or changes in the grade assignments of teachers is crucial for successfully refining the incentive scheme to account for ratcheting behavior.[41]

In Appendix F.3 (available online), I use within-school teacher-grade assignments and pre-reform quality measures to show that the pattern of teacher re-allocation immediately after the reform is consistent with the dynamic gaming hypothesis. Building upon this direct evidence, I decompose the difference-in-differences and triple-differences effects to establish the comparative importance of the effort and sorting channels. In particular, I construct such effects using the subset of teacher-year observations corresponding to teachers who have not been reassigned to teach a new grade in any prior post-reform period ($\Delta g = 0$) and those who have in at least one prior post-reform period ($\Delta g \neq 0$).[42] In the former case, any dynamic gaming effects should exclusively be due to differential teacher effort, while effects for the latter case are expected to arise from a combination of differential effort and re-allocation.

Table 4 presents the results of the decomposition, both within each group ($\Delta g = 0$ and $\Delta g \neq 0$) using variation in group proportions across schools (the "Unadjusted for Share" effects) and across groups by conditioning on the share of each in the sample to produce effects which sum to the total effects for all teachers (the "Adjusted for Share" effects). Comparing K-5 to both K-8 and K-6 schools, the evidence is consistent with the effort and sorting mechanisms being important. For the comparison between K-5 and K-8 schools, the unadjusted triple-differences estimates are positive and significant for both groups and the point estimates are slightly larger for the $\Delta g \neq 0$ group (although not significantly so). Moreover, while neither of the positive triple-differences estimates are significant for the comparison between K-5 and K-6 schools (which is in keeping with earlier findings), the difference-in-differences estimates are both positive and significant, with a slightly larger point estimate found once again for the $\Delta g \neq 0$ group. This accords with intuition, as the $\Delta g \neq 0$ group reflects both dynamic gaming channels rather than only effort for $\Delta g = 0$.

[41] One might argue that the shifting of non-teacher-based resources across grades is a potentially important third channel. While I do not possess data on all such inputs, I am able to observe class size and can rule out changes in it from driving the main results: all difference-in-differences and triple-differences estimates using class size (rather than test scores) as the dependent variable are insignificant, though of the expected negative sign. These results are available upon request.

[42] I do not exploit pre-reform teacher quality measures for this decomposition, since they do not exist for a majority of post-reform teacher observations.

Table 4
Decomposing the Dynamic Effects

A. Estimates for K5-K8 Comparison

	Unadjusted for Share		Adjusted for Share	
	$\Phi_{5,post-pre}$	$\Phi_{5-4,post-pre}$	$\Phi_{5,post-pre}$	$\Phi_{5-4,post-pre}$
$\Delta g = 0$	1.14**	1.34*	0.60**	0.71*
	(0.53)	(0.71)	(0.28)	(0.37)
$\Delta g \neq 0$	1.13*	1.49*	0.53*	0.70*
	(0.63)	(0.87)	(0.30)	(0.41)
Total Effect	—	—	1.13**	1.41**
			(0.46)	(0.55)

B. Estimates for K5-K6 Comparison

	Unadjusted for Share		Adjusted for Share	
	$\Phi_{5,post-pre}$	$\Phi_{5-4,post-pre}$	$\Phi_{5,post-pre}$	$\Phi_{5-4,post-pre}$
$\Delta g = 0$	1.42**	1.05	0.76**	0.56
	(0.67)	(0.74)	(0.36)	(0.39)
$\Delta g \neq 0$	1.94**	0.96	0.91**	0.45
	(0.78)	(0.76)	(0.36)	(0.35)
Total Effect	—	—	1.67***	1.01
			(0.61)	(0.62)

NOTE.—This table presents difference-in-differences and triple-differences estimates for teachers who have not been reassigned to teach a new grade in any prior post-reform period ($\Delta g = 0$) and those who have in at least one prior post-reform period ($\Delta g \neq 0$). To facilitate such an analysis, only teachers whose work history spans the years 1997 through the year of observation are included. The estimates are constructed from joint F-tests of the interaction dummies (pre/post period × type × grade × grade change classification) included in the regression with full controls (specification (3)) for the subsample of schools that do not switch configuration during the period of analysis. Owing to the institutional details of the reform, the pre- and post-reform period is respectively defined as 1995 through 1996 and 1998 through 2005. Both unadjusted effects that indicate the magnitude of dynamic gaming within each group of teachers and effects that are adjusted for the share of each group in the overall sample are reported. The latter estimates sum to the effect across both groups, which is reported in the third row. Standard errors adjusted for clustering at the school level are reported in parentheses.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Under the more restrictive pre- and post-reform definitions, the difference-in-differences and triple-differences estimates across all teachers (found in the "Total Effect" row of each panel) are similar to those for the full sample. Focusing on the preferred triple-differences results which are significant for the comparison between K-5 and K-8 schools, the adjusted-for-share estimates reveal that each group accounts for half of the total dynamic gaming effect. Given that the effort channel is expected to account for the entire $\Delta g = 0$ estimate and a portion of the $\Delta g \neq 0$ estimate, this suggests that differential effort is a key driver of dynamic gaming and potentially the primary channel through which it occurs, while teacher re-allocation is likely an important secondary channel.

VII. Conclusion

A broad class of incentive schemes in education and elsewhere condition on prior outcomes to compensate for heterogeneous inputs. While increased efficiency is likely to result from such a design, these schemes make it possible for agents to manipulate future targets by distorting contemporaneous effort decisions. Credible empirical estimates of these distortions are scarce and no previous studies have explored this issue in an education context, where conditioning accountability targets on prior performance has become increasingly prevalent.

A primary reason for this state of affairs is that existing theoretical analyses do not provide a clear prediction as to where one might look for such dynamic effects – an important element in forming a plausible identification strategy. In this paper, I develop a novel test for these distortions in an educational setting by reformulating the prior dynamic moral hazard theory to accommodate ratchet effects with finite horizons and human capital accumulation. These extensions produce a viable research design where ratchet effects are identified from variation in the horizons schools face, as captured by the school grade span.

Using a triple-differences strategy to account for differentially trending unobservables across schools, I find substantial evidence of such effects, with distortions ranging between 3.9 and 5.9 percent of a standard deviation in the grade five score. Several robustness checks lend credence to my dynamic gaming interpretation of the results. Exploiting additional data on teacher-grade assignments, I also provide insight into the mechanisms that generate the estimated effects. The evidence indicates that they are likely to be driven primarily by distortions in classroom effort, with re-sorting of teachers across grades serving as an important secondary channel.

Given the substantial stakes often associated with incentive schemes in education and beyond, it is important that policymakers are cognizant of the nontrivial distortions that can arise when future targets are manipulable. I propose an alternative target that eliminates these distortions by sacrificing a portion of the efficiency gained through conditioning on prior outcomes. This provides a foundation for designing a more refined scheme that achieves an optimal balance between accounting for variation in inputs and limiting dynamic gaming, a subject I plan to pursue in future work.

References

Allen, Douglas W., and Dean Lueck. 1999. Searching for ratchet effects in agricultural contracts. *Journal of Agricultural and Resource Economics* 24, no. 2: 536–552.

Alspaugh, John W. 1998. Achievement loss associated with the transition to middle school and high school. *Journal of Educational Research* 92, no. 1: 20–25.

Baron, David P., and David Besanko. 1987. Commitment and fairness in a dynamic regulatory relationship. *Review of Economic Studies* 54, no. 3: 413–436.

Bifulco, Robert, and Helen F. Ladd. 2006. The impacts of charter schools on student achievement: Evidence from North Carolina. *Education Finance and Policy* 1, no. 1: 50–90.

Carnoy, Martin, and Susanna Loeb. 2002. Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis* 24, no. 4: 305–331.

Charness, Gary, Peter Kuhn, and Marie-Claire Villeval. 2010. Competition and the ratchet effect. Working Paper no. 16325, National Bureau of Economic Research, Cambridge, MA.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2013a. Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. Working Paper no. 19423, National Bureau of Economic Research, Cambridge, MA.

——. 2013b. Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. Working Paper no. 19424, National Bureau of Economic Research, Cambridge, MA.

Cooper, David J., John H. Kagel, Wei Lo, and Qing Liang Gu. 1999. Gaming against managers in incentive systems: Experimental results with Chinese students and Chinese managers. *American Economic Review* 89, no. 4: 781–804.

Cullen, Julie B., and Randall Reback. 2006. Tinkering toward accolades: School gaming under a performance accountability system. Working Paper no. 12286, National Bureau of Economic Research, Cambridge, MA.

Dee, Thomas S., and Brian Jacob. 2011. The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management* 30, no. 3: 418–446.

Fabrizio, Louis M. 2006. The creation and evolution of North Carolina's ABCs accountability program and the impact of No Child Left Behind - A case study. PhD diss., N.C. State University.

Figlio, David N., and Lawrence W. Kenny. 2007. Individual teacher incentives and student performance. *Journal of Public Economics* 91, no. 5: 901–914.

Freixas, Xavier, Roger Guesnerie, and Jean Tirole. 1985. Planning under incomplete information and the ratchet effect. *Review of Economic Studies* 52, no. 2: 173–191.

Fruehwirth, Jane Cooley. 2013. Identifying peer achievement spillovers: Implications for desegregation and the achievement gap. *Quantitative Economics* 4, no. 1: 85–124.

Gibbons, Robert. 1987. Piece-rate incentive schemes. *Journal of Labor Economics* 5, no. 4: 413–429.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. 2004. Disruption versus tiebout improvement: The costs and benefits of switching schools. *Journal of Public Economics* 88, no. 9: 1721–1746.

Hanushek, Eric A., and Margaret E. Raymond. 2005. Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management* 24, no. 2: 297–327.

Heneman, Herbert G. 1998. Assessment of the motivational reactions of teachers to a school-based performance award program. *Journal of Personnel Evaluation in Education* 12, no. 1: 43–59.

Holmstrom, Bengt. 1982. Design of incentive schemes and the new Soviet incentive model. *European Economic Review* 17, no. 2: 127–148.

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, no. 3: 843–877.

Juvonen, Jaana, Vi-Nhuan Le, Tessa Kaganoff, Catherine H. Augustine, and Louay Constant. 2004. *Focus on the wonder years: Challenges facing the american middle school.* Santa Monica, CA: RAND Corporation.

Kane, Thomas J., and Douglas O. Staiger. 2001. Improving school accountability measures. Working Paper no. 8156, National Bureau of Economic Research, Cambridge, MA.

——. 2002. The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16, no. 4: 91–114.

——. 2008. Estimating teacher impacts on student achievement: An experimental evaluation. Working Paper no. 14607, National Bureau of Economic Research, Cambridge, MA.

Kanemoto, Yoshitsugu, and W. Bentley MacLeod. 1992. The ratchet effect and the market for secondhand workers. *Journal of Labor Economics* 10, no. 1: 85–98.

Keren, Michael, Jeffrey Miller, and James R. Thornton. 1983. The ratchet: A dynamic managerial incentive model of the Soviet enterprise. *Journal of Comparative Economics* 7, no. 4: 347–367.

Ladd, Helen F. 2001. School-based educational accountability systems: The promise and the pitfalls. *National Tax Journal* 54, no. 2: 385–400.

Ladd, Helen F., and Arnaldo Zelli. 2002. School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly* 38, no. 4: 494–529.

Laffont, Jean-Jacques, and Jean Tirole. 1988. The dynamics of incentive contracts. *Econometrica* 56, no. 5: 1153–1175.

Lavy, Victor. 2002. Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of Political Economy* 110, no. 6: 1286–1317.

——. 2009. Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review* 99, no. 5: 1979–2011.

Lazear, Edward P. 1986. Salaries and piece rates. *Journal of Business* 59, no. 3: 405–431.

McCaffrey, Daniel F, J. R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton. 2004. Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics* 29, no. 1: 67–101.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. Teacher performance pay: Experimental evidence from India. *Journal of Political Economy* 119, no. 1: 39–77.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics* 92, no. 2: 263–283.

Parent, Daniel. 1999. Methods of pay and earnings: A longitudinal analysis. *Industrial and Labor Relations Review* 53, no. 1: 71–86.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73, no. 2: 417–458.

Rockoff, Jonah E., and Benjamin B. Lockwood. 2010. Stuck in the middle: Impacts of grade configuration in public schools. *Journal of Public Economics* 94, no. 11: 1051–1061.

Rothstein, Jesse. 2010. Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics* 125, no. 1: 175–214.

Todd, Petra E., and Kenneth I. Wolpin. 2003. On the specification and estimation of the production function for cognitive achievement. *Economic Journal* 113, no. 485: F3–F33.

——. 2007. The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human capital* 1, no. 1: 91–136.

Weitzman, Martin L. 1980. The ratchet principle and performance incentives. *Bell Journal of Economics* 11, no. 1: 302–308.

Appendices

A. Dynamic Moral Hazard Literature

A.1. Infinite Horizon Without Production Technology

The model in Section II has features in common with several existing papers, but it is most easily motivated by building upon the seminal work of Weitzman (1980), which predicts the emergence of ratchet effects when performance today determines bonus receipt today and tomorrow. In Weitzman's model, a fixed linear incentive scheme rewards agents based on the difference between a current output measure y_t and the target αy_{t-1}, which is an adjusted prior measure. The adjustment parameter α dictates how much the principal (in the 'principal-agent' sense) must reward agents, conditional on current and prior output as well as the linear payment scheme reward parameter b. To see this, consider an agent's problem at time t: Given the scheme and a convex cost of output $C(\cdot)$, the agent's objective is given by

$$\max_{\{y_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \delta^t [b(y_t - \alpha y_{t-1}) - C(y_t)]$$

which leads to the first-order conditions $b(1 - \delta\alpha) = C'(y_t), \forall\, t$. Comparing this to the condition without dynamic considerations, $b = C'(y_t), \forall\, t$, which occurs if the target is α instead of αy_{t-1}, the ratchet effect leads workers to underperform if $\delta\alpha > 0$.[43] Intuitively, as α increases, the next period target rises when contemporaneous output is unchanged, which results in lower pay in the following period. Therefore, the marginal benefit of output decreases as α increases, which results in a lower optimal level of output, given the same marginal cost. This effect is magnified as future periods are discounted less by the agent (higher δ).

[43] By definition, the inter-temporal depreciation rate δ is positive, while the target α will also be positive if it is derived by regressing a current positive measure on a smaller positive prior one (as in my empirical application, for example).

B. The North Carolina Accountability Policy

B.1. Background

Stemming from legislation ratified in June 1996 and implemented in 1997 (the 1996-97 school year),[44] North Carolina's high-stakes accountability system consists of monetary rewards for outperformance of a composite school-level growth target.[45] Performance is measured by norm-referenced end-of-grade reading and mathematics tests taken by students in grades three through eight, beginning in 1993. Under the scheme, if the difference between the composite realized growth Δy_{st} and expected growth target $\Delta \widehat{y}_{st}$ for school s in year t is positive, then the principal and all teachers at the school each receive additional compensation of \$750; otherwise, they do not. If the school exceeds a further target that is set 10 percent higher than the expected growth target, then the bonus is increased to \$1,500.[46]

It is worth briefly elaborating on how the composite realized growth and target growth for a school are calculated. Both the school-level realized growth Δy_{st} and expected growth target $\Delta \widehat{y}_{st}$ are composites of their respective subject- and grade-specific counterparts. The composite realized growth for each school is equal to the mean of all pertinent subject-grade scores for that school (each one itself averaged across the students in that subject-grade pair), weighted according to the historical standard deviation of all scores for each subject-grade in the state (σ^z_{gt}, where $z \in \{r, m\}$ denotes the subject).[47] The composite school-level

[44]Ratification and implementation occurred following a pilot phase in 1996 covering ten districts containing 63 schools (about 4 percent of all primary and middle schools in North Carolina), and after a similar accountability system was put in place in the Charlotte-Mecklenburg district for the same school year. Both of these were supplanted by the state-wide reform, which was expanded in 1998 to encompass high school students as well. Prior to 1996, and for all other schools in 1996, a low-stakes accountability environment was in place, consisting of published district and then school report cards (see Fabrizio (2006) and Ladd (2001) for greater detail on the historical underpinnings of state testing and reporting in North Carolina). Basic information about the North Carolina ABCs of Public Education (which stands for strong Accountability, teaching the Basics and focusing on local Control) is found in an electronic brochure at *http://www.ncpublicschools.org/docs/accountability/reporting/abc/2005-06/abcsbrochure.pdf* and a copy of a more detailed timeline is available in Appendix A of Fabrizio (2006).

[45]In general, accountability schemes tend to be implemented at the school level. This may be motivated from an incentive design standpoint, given that the yearly variation in transitory processes that Kane and Staiger (2002) highlight will be magnified when scores are averaged across a smaller group of students.

[46]A teacher with 13 years of experience and a bachelor's degree made about \$30,000 in 1998. Thus, \$1,500 is approximately equal to 5 percent of yearly pay or 60 percent of monthly pay.

[47]The standard deviations are calculated using score data in 1995 and these fixed values are used in all future years through 2005.

growth target is aggregated in an analogous way, using pertinent subject-grade targets with the same weightings.

The underlying grade-specific expected growth targets for reading and mathematics are calculated for each student using her prior performance according to the following formulae:

$$\Delta\widehat{r}_{igst} = \hat{\alpha}_0^g + \hat{\alpha}_1^g(r_{isg-1t-1} - \bar{r}_{g-1t-1} + m_{isg-1t-1} - \bar{m}_{g-1t-1}) + \hat{\alpha}_2^g(r_{isg-1t-1} - \bar{r}_{g-1t-1})$$

$$\Delta\widehat{m}_{igst} = \hat{\beta}_0^g + \hat{\beta}_1^g(r_{isg-1t-1} - \bar{r}_{g-1t-1} + m_{isg-1t-1} - \bar{m}_{g-1t-1}) + \hat{\beta}_2^g(m_{isg-1t-1} - \bar{m}_{g-1t-1})$$

where $\Delta\widehat{r}_{igst} \equiv \widehat{r}_{igst} - r_{isg-1t-1}$, $\Delta\widehat{m}_{igst} \equiv \widehat{m}_{igst} - m_{isg-1t-1}$; r_{igst} and m_{igst} are the average reading and math scores for student i in school s, grade g and year t; $\bar{r}_{gt}$ and $\bar{m}_{gt}$ are the average reading and math scores across all schools in the state for grade g in year t;[48] and the grade-specific coefficients $\hat{\alpha}_0^g$, $\hat{\alpha}_1^g$, $\hat{\alpha}_2^g$, $\hat{\beta}_0^g$, $\hat{\beta}_1^g$ and $\hat{\beta}_2^g$ are given. The coefficients are estimated from score data in 1993 and 1994 by regressing the actual score gain in 1994 for reading or mathematics on the lagged reading and mathematics scores according to the preceding formulae and are fixed at these values for all subsequent years through 2005.[49] Using lagged school-specific scores along with the fixed coefficients and state average lagged scores, the expected growth targets (or gains) are calculated for every grade in a school for each year beginning in 1997.

The first component of each expected gain ($\hat{\alpha}_0^g$ or $\hat{\beta}_0^g$) is the mean expected gain across all schools in the state for grade g. The second component is the sum of the demeaned prior performance in both subjects and is treated as a proxy for average student ability in the school. The third component is the demeaned prior performance in the subject for which the expected gain is being calculated, and is used as a correction for mean reversion. Consider, for instance, schools that had above-average scores in both reading and math; they would be expected to outperform an average school due to having a more able student body, but their expected performance would be attenuated by the tendency for atypical scores to correct

[48] As with the statewide subject-grade standard deviations, the state means ($\bar{r}_{g-1t-1}$ and $\bar{m}_{g-1t-1}$) are fixed over time using the score data in 1995.

[49] Specifically, using $t = 1994$, $\hat{\alpha}_0$, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are obtained from the first equation, while $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are obtained from the second one. In 2001, these reduced-form coefficients were then updated for grade three only. I have verified that this recipe produces the coefficients used by the North Carolina accountability scheme by independently implementing it. I also extended the analysis to all pre- and post-reform years, finding that the reduced-form targets are highly dependent on the reference year that is selected.

toward the state average over time.[50]

In essence, the North Carolina incentive scheme uses one year of prior school performance to proxy for all prior inputs. It also attempts to exploit the disparity between reading and math scores to control for any component of the prior score that does not contribute permanently to a child's learning in the future. Given the structure of the North Carolina approach, there are several reasons why targets may be too easy or difficult to satisfy, stemming from the fact that the combined prior reading and math scores are not exclusively the result of student ability. These include differences between prior and current teacher ability and/or school resources, and any transitory effect that is misinterpreted as a permanent one since it influences both subjects. They are undesirable aspects of the reform, since teachers are then held accountable for an outcome that they do not fully control. That said, my econometric strategy for uncovering ratcheting behavior does not depend on the existence of such contemporaneous inefficiencies.

While the 1996 North Carolina accountability reform is not without its flaws, there are several elements that make it well-suited for detecting evidence of dynamic gaming behavior. First, it features a high-powered school-level reward scheme that conditions targets on student prior scores to account for heterogeneity in students, teachers and resources, all of which are predicted to be key ingredients in generating ratchet effects within schools. Survey evidence lends credence to this idea.[51] Moreover, the program is long-standing, having dispensed over $870 million in pecuniary payments to educators through 2005 (Fabrizio 2006),[52] so that any dynamic distortions would have had time to manifest themselves.

[50]Kane and Staiger (2002) highlight the importance of year-to-year transitory shocks in determining scores. Ideally, an incentive scheme would not hold teachers accountable for factors that were out of their control. It is therefore notable that North Carolina policymakers made an effort to correct for mean-reverting processes.

[51]Referring to the 1995-96 scheme in Charlotte-Mecklenburg that has strong similarities to the North Carolina accountability program that followed, Heneman (1998) reports that very few teachers agreed with the statement: "We can continue to meet ever-higher student achievement goals in the future." This suggests that they were thinking about dynamic consequences when the program was introduced.

[52]Only minor changes were made to the program prior to a more substantial overhaul in 2006, which explains why my investigation into dynamic gaming does not extend beyond 2005. Despite the changes to the reform in 2006, the salient features that are necessary for such gaming to occur remain in place, as the targets continue to be at the school level and depend on student prior scores.

C. Proofs

C.1. Proof of Lemma 1

Using the preceding conditions, observe that the difference between effort levels in consecutive grades is $e_{scgt} - e_{scg-1t} = \frac{b\delta}{d}(\alpha - \gamma)\delta^{G_c-g}\gamma^{G_c-g}f_c(\Pi_{sct+1+G_c-g})$ for $2 \leq g \leq G_c$. $f_c(\cdot) \geq 0$ then implies that $e_{scgt} \geq e_{scg-1t}$ for $2 \leq g \leq G_c$ (with a strict inequality holding if $f_c(\Pi_{sct+1+G_c-g}) > 0$).

C.2. Proof of Proposition 1

For some positive integer κ, consider arbitrary grade structures, with $G_c = G$ and $G_{c'} = G+\kappa > G_c$. Let us first compare the effort choices between these two types for grade $g \in \mathcal{G}_c$, where $\mathcal{G}_c$ is the set of grades served by a school with configuration c.

If $g = G$, then $e_{cG} = \frac{b}{d}$ and $e_{c'G} = \frac{b}{d}\left[1 + \delta(\gamma - \alpha)\sum_{i=0}^{\kappa-1}\delta^i\gamma^i\right]$, which means that $e_{cG} > e_{c'G}$ from the stated assumptions. If $1 \leq g < G$, then $e_{cg} = \frac{b}{d}\left[1 + \delta(\gamma - \alpha)\sum_{i=0}^{G-g-1}\delta^i\gamma^i\right]$ and $e_{c'g} = \frac{b}{d}\left[1+\delta(\gamma-\alpha)\sum_{i=0}^{G+\kappa-g-1}\delta^i\gamma^i\right]$. Since $\sum_{i=0}^{G+\kappa-g-1}\delta^i\gamma^i = \sum_{i=0}^{G-g-1}\delta^i\gamma^i + \sum_{i=G-g}^{G+\kappa-g-1}\delta^i\gamma^i > \sum_{i=0}^{G-g-1}\delta^i\gamma^i$, using the stated assumptions, we have $e_{cg} > e_{c'g}$. If $g = 0$, then $e_{cg} = \frac{b}{d}\left[\delta(\gamma - \alpha)\sum_{i=0}^{G-1}\delta^i\gamma^i\right]$ and $e_{c'g} = \frac{b}{d}\left[\delta(\gamma - \alpha)\sum_{i=0}^{G+\kappa-1}\delta^i\gamma^i\right]$, given that there is no contemporaneous benefit to exerting effort in the untested grade $g = 0$. Since $\sum_{i=0}^{G+\kappa-1}\delta^i\gamma^i = \sum_{i=0}^{G-1}\delta^i\gamma^i + \sum_{i=G}^{G+\kappa-1}\delta^i\gamma^i > \sum_{i=0}^{G-1}\delta^i\gamma^i$, using the stated assumptions, we have $e_{c0} > e_{c'0}$. Therefore, $e_{cg} > e_{c'g}, \forall\, g \in \mathcal{G}_c$.

Given the assumptions about initial educational capital stock and teacher ability, let $k_{c0} = k_0, \forall\, c$ (where k_{c0} is defined to be the capital stock in grade $g = 1$), and $a_{cg} = a_g, \forall\, c$. Assuming that the shock at the average school of each type c is zero ($u_{cg} = 0$), the test score for any type c school is $y_{cg} = \gamma^{g+1}k_0 + \sum_{i=0}^{g}\gamma^{g-i}a_i + \sum_{i=1}^{g}\gamma^{g-i}e_{ci}$.

Since $e_{cg} > e_{c'g}, \forall\, g \in \mathcal{G}_c$, it should be immediate from the preceding expression that $y_{cg} > y_{c'g}, \forall\, g \in \mathcal{G}_c$, which is the desired result.

C.3. Proof of Proposition 2

Recall from the proof of Proposition 1 that $\sum_{i=0}^{G+\kappa-g-1}\delta^i\gamma^i = \sum_{i=0}^{G-g-1}\delta^i\gamma^i + \rho_{\kappa g}$, where $\rho_{\kappa g} \equiv \sum_{i=G-g}^{G+\kappa-g-1}\delta^i\gamma^i$. If $\delta\gamma < 1$, then $\rho_{\kappa g}$ is increasing in g, since each term in the sum is less

than one and is raised to a power that is decreasing in g. Thus, $\sum_{i=0}^{G+\kappa-g-1} \delta^i \gamma^i - \sum_{i=0}^{G-g-1} \delta^i \gamma^i$ is increasing in g, which means that $e_{cg} - e_{c'g}$ is increasing in g, $\forall\, g \in \mathcal{G}_c$. Therefore, under the same assumptions of Proposition 1, $y_{cg} - y_{c'g}$ is increasing in g, $\forall\, g \in \mathcal{G}_c$.

D. Data

D.1. Available Data

The student-level data extends from 1993 to 2011.[53] However, the reform was substantially altered in 2006 (as reflected by the double horizontal separator in the graphical representation of the data by year and cohort below) and data for 1993 cannot be linked with later years. Data for 1996 are also missing for grades five through eight, but I am able to overcome this limitation for grades five through seven in 1996 by using the prior year scores for grades six through eight in 1997. School-level characteristics are then imputed for grade five students in 1996 who attend a 6-8 middle school in 1997, by constructing composite K-5 feeder schools from the K-5 schools that feed each 6-8 school in 1998.

	Cohort															
Year	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1993	4	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1994	5	4	3	-	-	-	-	-	-	-	-	-	-	-	-	-
1995	6	5	4	3	-	-	-	-	-	-	-	-	-	-	-	-
1996	-	-	-	4	3	2	-	-	-	-	-	-	-	-	-	-
1997	-	-	6	5	4	3	2	-	-	-	-	-	-	-	-	-
1998	-	-	7	6	5	4	3	2	-	-	-	-	-	-	-	-
1999	-	-	8	7	6	5	4	3	2	-	-	-	-	-	-	-
2000	-	-	-	8	7	6	5	4	3	2	-	-	-	-	-	-
2001	-	-	-	-	8	7	6	5	4	3	2	-	-	-	-	-
2002	-	-	-	-	-	8	7	6	5	4	3	2	-	-	-	-
2003	-	-	-	-	-	-	8	7	6	5	4	3	2	-	-	-
2004	-	-	-	-	-	-	-	8	7	6	5	4	3	2	-	-
2005	-	-	-	-	-	-	-	-	8	7	6	5	4	3	2	-
2006	-	-	-	-	-	-	-	-	-	8	7	6	5	4	3	2
2007	-	-	-	-	-	-	-	-	-	-	8	7	6	5	4	3
2008	-	-	-	-	-	-	-	-	-	-	-	8	7	6	5	4
2009	-	-	-	-	-	-	-	-	-	-	-	-	8	7	6	5
2010	-	-	-	-	-	-	-	-	-	-	-	-	-	8	7	6
2011	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8	7

[53] With respect to score data, although second editions of the tests for elementary students in mathematics and reading were introduced in 2001 and 2003, respectively, state psychometricians use equating studies to not only ensure comparability across grades but years as well.

E. The Impact of the Reform in Pictures

It is instructive to see which patterns emerge in the data. There are three features that are particularly interesting. The first relates to whether the reform had a positive effect on scores overall. Evidence presented in Figure A.3 suggests that it did, which is in keeping with the objective of policymakers. It shows density plots of first-differenced student scores by grade, for grades two through five, using scores that are adjusted for observable characteristics.[54] The mean of each distribution is positive, reflecting the fact that the average post-reform score is greater than its pre-reform counterpart.

The second notable feature of the data apparent in Figure A.3 is that the growth in scores is monotonically increasing in the grade, which is the type of dynamic pattern predicted by the theoretical model. Moreover, growth in the average grade two score is nearly zero and is certainly much lower than is observed for the higher grades. Although it is not a focus of my econometric strategy, the model would predict that the effort in this untested grade should be as low as possible to reduce the target for grade three, given that there is no contemporaneous benefit of exerting effort in grade two. The corresponding distribution is consistent with this prediction.

The third interesting feature is that the effect of the reform was not uniform across school configurations. Figure A.5 decomposes the grade five score by school configuration, plotting the density and means (given by the vertical lines) of the first-differenced grade five score for K-5, K-6 and K-8 schools, respectively. Recall from Proposition 1 that, controlling for differences in the initial educational capital of students and teacher ability, the school with a shorter grade horizon will have a higher test score than one with a longer horizon. Using the pre-reform period as a baseline and conditioning on student and school characteristics, the figure reveals evidence consistent with this proposition. In particular, the mean for K-5 schools is higher than the mean for either K-6 or K-8 schools, which are the main comparisons of interest (there are an insufficient number of observations to compare K-6 and K-8 schools).

[54] Across school-grade pairs, the average score in the pre- and post-reform period is regressed on controls, such as parental education and ethnicity, and the difference between the residuals before and after the reform is computed for each pair. Density plots are then formed for each grade using these differences. A similar pattern holds using test scores without controls (see Figure A.4).

F. Discussion of Additional Results

F.1. Evolution of Dynamic Effects

Exploiting the abundance of post-reform data, Table A.5 reports the difference-in-differences and triple-differences estimates for three post-reform periods of equal duration. The evidence indicates that the disparity evolves as one might expect if principals and teachers initially require time to acclimate to the new incentive environment. In particular, comparing either K-5 and K-8, or K-5 and K-6 schools, the estimates for 1997-1999 are smaller than the analogous ones for 2000-2002, and statistically so in the difference-in-differences case.

F.2. Strength of Dynamic Effects

Whether teachers respond to the school-level incentives in a decentralized way or their effort is centrally coordinated through the principal, it would be reasonable to expect that the dynamic gaming effect would be more pronounced for schools with fewer teachers per grade.[55] On the other hand, while a greater number of classes might lead to more free riding, such a situation would also allow greater flexibility for a principal engaging in teacher re-allocation across grades. Thus, as the number of classes increases, the dynamic gaming effect is expected to be attenuated under an effort-based channel and strengthened under a sorting-based channel. On balance, the evidence lends support for the former channel. Dividing the difference-in-differences and triple-differences effects according to schools with a small and large number of tested classes per grade, Table A.7 shows that the point estimates are larger for the smaller classification in all cases and significantly so for the difference-in-differences estimates.

School-specific principal tenure may also potentially affect the extent of dynamic gaming that occurs through the coordination of effort and re-sorting of teachers by principals. Decomposing the difference-in-differences and triple-differences effects according to whether the school principal is new to the school or more established, Table A.8 reveals that the point estimates for principals with more than one year of school-specific experience are larger than

[55] In the former case, the free rider problem results in an effect that diminishes as the number of teachers in a school increases. In the centralized latter case, the principal might find it more difficult to coordinate over a greater number of teachers.

for those without school-specific experience. Although this difference is not statistically significant, owing to the large standard errors, this evidence is at least suggestive that the dynamic gaming response by school principals is greater when their information set about teachers is more complete.

F.3. Direct Evidence of Teacher Re-allocation

Within-school teacher re-allocation stands apart from differential effort in that it is directly observed in the data. This allows for an analysis of teacher re-allocation by quality, to determine if the re-sorting that occurs is consistent with the dynamic gaming hypothesis. Such an investigation depends on the existence of a reliable measure of teacher quality which can be matched to a high proportion of teachers in the post-reform period. As 1998 is the first year in which school principals could have plausibly dynamically gamed the system through teacher re-allocation, I compute fixed effects for each teacher in 1997 and use them to analyze teacher re-sorting in 1998 only.[56]

Table A.10 reports the results of the re-allocation analysis by teacher quality. Defining a high (h) and low (l) quality teacher as possessing a fixed effect that is above and below the median, respectively, I compute the change in assigned grade (Δg) from 1997 to 1998 for each group and the difference between them.[57] I perform this calculation for the entire sample of teachers and a variety of teacher experience cutoffs. Across all teachers, there is statistically no difference between the re-allocation of high and low quality teachers. However, in line with the idea that more entrenched teachers may be more likely to resist re-sorting, striking results emerge when limiting the analysis to teachers with fewer than thirteen years of experience. Relative to their low quality counterparts, high quality teachers are re-sorted up the grade distribution and the effect is statistically significant: of those teachers who switch grades,

[56]Given that the accountability reform was introduced in June of 1996 after allocation decisions had been made for 1997, it is unlikely that teacher re-sorting for dynamic gaming purposes occurred prior to 1998. Although using fixed effects from 1997 is not ideal, as they may reflect some early dynamic distortions of effort, it is a reasonable approximation that maximizes the number of teachers for whom teaching quality and re-allocation can jointly be observed (note that score data is missing for grades five through eight in 1996 and the match rate using analogous 1995 quality measures or later post-reform years is substantially lower).

[57]For example, Δg would equal 1 if a teacher was reassigned to grade five from four (or four from three) and 0 if not reassigned. Other transitions are calculated analogously.

the average effect is an increase of about three tenths of a grade. Decomposing the relatively inexperienced teachers into four three-year intervals of experience reveals that the results are driven by teachers who have taught for one to three years and seven to nine years: in each case, reassigned high quality teachers are significantly moved up by nearly half a grade when compared to their low quality counterparts. The lack of a significant result for teachers who have taught for four to six years is particularly interesting, given that tenure is granted in the fourth year of teaching in North Carolina and tenure could make teachers relatively more difficult to re-sort as they gain bargaining power.

Table A.1
Regression Specifications

Dependent Variable: Combined Mathematics and Reading Score			
Specification:	(1)	(2)	(3)
Student - Parental Education:			
No High School		-17.36***	-17.29***
		(0.15)	(0.16)
High School Graduate		-10.55***	-10.45***
		(0.13)	(0.14)
Trade School		-5.78***	-5.60***
		(0.13)	(0.13)
Community College		-6.62***	-6.55***
		(0.13)	(0.13)
4-Year College		-2.51***	-2.45***
		(0.10)	(0.10)
Student - Ethnic - Black		-8.53***	-8.44***
		(0.10)	(0.10)
Student - Exceptionality:			
Learning Impairment		-13.23***	-13.22***
		(0.09)	(0.10)
Gifted/Exceptional		16.65***	16.71***
		(0.09)	(0.09)
Prop. Free Lunch Eligible		-5.24***	-4.83***
		(0.35)	(0.36)
Teacher Licensure Test Score			0.62***
			(0.06)
Constant	283.3***	349.1***	348.9***
	(0.8)	(0.2)	(0.2)
School Locale Controls?	No	Yes	Yes
R^2	0.338	0.664	0.670
Observations	6,130,308	5,318,520	4,499,997

Note: This table defines three specifications according to the components included in the control vector of the main estimating equation (equation (3)) and reports the coefficient for each component. All specifications include interaction dummies (pre/post period × type × grade), which are used to construct the first-differences, difference-in-differences and triple-differences estimates reported in Table 2. The analysis is conducted for the years 1994 through 2005, with the number of observations declining as regressors with missing values are added. Standard errors adjusted for clustering at the school level are reported in parentheses.

*** Significant at the 1 percent level.

Table A.2
Evidence Against Household Sorting

	K5 vs. K8		K5 vs. K6	
Estimate:	$\Phi_{5,post-pre}$	$\Phi_{5-4,post-pre}$	$\Phi_{5,post-pre}$	$\Phi_{5-4,post-pre}$
Parental Education:				
No High School	0.009	0.000	-0.012	-0.021
	(0.010)	(0.012)	(0.014)	(0.017)
High School Grad	-0.007	-0.023	0.039**	0.016
	(0.013)	(0.016)	(0.018)	(0.020)
Trade School	-0.003	0.006	0.005	0.005
	(0.005)	(0.008)	(0.007)	(0.013)
Community College	-0.013	0.003	-0.023*	0.004
	(0.009)	(0.011)	(0.012)	(0.013)
4-Year College	0.013	0.012	-0.007	-0.005
	(0.009)	(0.011)	(0.011)	(0.011)
Graduate Degree	0.001	0.001	-0.003	0.001
	(0.005)	(0.006)	(0.005)	(0.005)
Prop. Free Lunch	0.036***	-0.009	0.037**	0.018
	(0.011)	(0.008)	(0.016)	(0.015)

Note: To rule out household sorting based on observable characteristics, this table presents difference-in-differences and triple-differences estimates for parental education and free lunch status. The estimates are constructed from joint F-tests of the interaction dummies included in a regression for the subsample of schools that do not switch configuration during the period of analysis. Standard errors adjusted for clustering at the school level are reported in parentheses.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Table A.3
Results by School Locale

Estimate:	$\Phi_{K5,5-4,post-pre}$	$\Phi_{K5-K8,5-4,post-pre}$
Locale		
All	1.19***	1.66***
	(0.17)	(0.51)
Rural	1.34***	2.22***
	(0.29)	(0.58)
Urban	1.22***	—
	(0.27)	
Suburban	0.84***	—
	(0.31)	

Note: This table presents difference-in-differences estimates within K-5 schools by school locale (in contrast to the usual cross-configuration comparison, the relevant dimensions here are pre/post and grade). The familiar triple-differences estimates between K-5 and K-8 schools are also reported if there exists a sufficient number of K-8 schools (given the concentration of K-8 schools in rural areas, these results are therefore reported for "All" and "Rural." Each estimate is constructed from joint F-tests of the interaction dummies included in the regression with full controls (specification 3) for the subsample of schools that do not switch configuration during the period of analysis. Standard errors adjusted for clustering at the school level are reported in parentheses.

*** Significant at the 1 percent level.

Table A.4
Falsification Exercise

	K5 vs. K8		K5 vs. K6	
Estimate:	$\Phi_{5,post-pre}$	$\Phi_{5-4,post-pre}$	$\Phi_{5,post-pre}$	$\Phi_{5-4,post-pre}$
Year of Reform				
1996	0.91**	1.08**	1.05*	0.24
	(0.39)	(0.52)	(0.57)	(0.60)
1997	**1.36*****	**1.66*****	**1.51*****	**0.83**
	(0.38)	(0.51)	(0.56)	(0.59)
1998	1.03***	0.99**	1.18**	0.16
	(0.37)	(0.50)	(0.56)	(0.58)
1999	0.55	0.82	0.70	-0.01
	(0.37)	(0.50)	(0.56)	(0.58)
2000	0.06	0.42	0.21	-0.41
	(0.37)	(0.50)	(0.56)	(0.58)
2001	-0.23	0.03	-0.08	-0.80
	(0.37)	(0.50)	(0.56)	(0.58)
2002	-0.25	-0.41	-0.11	-1.25**
	(0.38)	(0.50)	(0.56)	(0.58)
2003	-0.21	-1.11**	-0.06	-1.95***
	(0.38)	(0.50)	(0.56)	(0.58)
2004	-0.76**	-0.98**	-0.62	-1.81***
	(0.38)	(0.50)	(0.56)	(0.58)
2005	-1.82***	-0.85*	-1.67***	-1.69***
	(0.38)	(0.52)	(0.57)	(0.60)

Note: This table presents the results of a falsification exercise where the reform is assumed to be first introduced in a year other than the actual one (1997 (1996-97 school year) in bold). For each counterfactual year (and the actual one), difference-in-differences and triple-differences estimates are reported, which are constructed from joint F-tests of the interaction dummies included in the regression with full controls (specification (3)) for the subsample of schools that do not switch configuration during the period of analysis. Standard errors adjusted for clustering at the school level are reported in parentheses.

*** Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

Table A.5
A Closer Look at the Post-Reform Period

	c = K5 vs. c' = K8			c = K5 vs. c' = K6		
Post-reform definition:	97-99	00-02	03-05	97-99	00-02	03-05
$\Phi_{c-c',5,post-pre}$	0.73*	1.66***	2.04***	0.68	2.40***	1.60**
	(0.41)	(0.46)	(0.46)	(0.44)	(0.85)	(0.71)
$\Phi_{c-c',5-4,post-pre}$	1.59***	1.93***	1.45**	1.08	1.19	0.17
	(0.52)	(0.57)	(0.63)	(0.65)	(0.74)	(0.71)

Note: This table reports difference-in-differences and triple-differences estimates constructed from joint F-tests of the interaction dummies included in the regression with full controls (specification (3)) for the subsample of schools that do not switch configuration during the period of analysis. The nine-year post-reform period is subdivided into three equal three-year periods when constructing the interaction dummies to analyze the evolution of the dynamic gaming effect. Standard errors adjusted for clustering at the school level are reported in parentheses.

*** Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

Table A.6
Supporting Evidence - Breakdown by Subject

	c = K5 vs. c' = K8		c = K5 vs. c' = K6	
Subject:	$M+R$	M	$M+R$	M
$\Phi_{c-c',5,post-pre}$	1.36***	1.02***	1.51***	1.13***
	(0.38)	(0.27)	(0.56)	(0.42)
$\Phi_{c-c',4,post-pre}$	-0.30	-0.33	0.69	0.57
	(0.48)	(0.30)	(0.69)	(0.46)
$\Phi_{c-c',3,post-pre}$	-0.51	-0.51	0.86	0.76
	(0.53)	(0.33)	(0.91)	(0.58)
$\Phi_{c-c',5-4,post-pre}$	1.66***	1.35***	0.83	0.56
	(0.51)	(0.33)	(0.59)	(0.39)

Note: This table compares difference-in-differences and triple-differences estimates for the combined score (mathematics and reading, or $M+R$) to those for mathematics (M). The estimates are constructed from joint F-tests of the interaction dummies included in the regression with full controls (specification (3)) for the subsample of schools that do not switch configuration during the period of analysis. The coefficient for reading is simply the difference between the Φ for $M+R$ and M. Standard errors adjusted for clustering at the school level are reported in parentheses.

*** Significant at the 1 percent level.

Table A.7
Coordination/Free-Riding Effects

	$c =$ K5 vs. $c' =$ K8			$c =$ K5 vs. $c' =$ K6		
Classes per grade:	S	L	$S-L$	S	L	$S-L$
$\Phi_{c-c',5,post-pre}$	2.06***	0.53	1.52*	2.16**	0.38	1.78*
	(0.62)	(0.69)	(0.92)	(0.84)	(0.61)	(0.96)
$\Phi_{c-c',5-4,post-pre}$	2.16***	1.74	0.42	1.27	0.06	1.21
	(0.75)	(1.59)	(1.86)	(0.85)	(1.19)	(1.57)

Note: This table presents difference-in-differences and triple-differences estimates for schools with a small (S) and large (L) number of classes per grade, and the difference between the two ($S-L$). The estimates are constructed from joint F-tests of the interaction dummies (pre/post period × type × grade × below/above 25th percentile of the number of classes per grade) included in the regression with full controls (specification (3)) for the subsample of schools that do not switch configuration during the period of analysis. A small number of classes per grade is defined to be three (the 25th percentile) or fewer. Standard errors adjusted for clustering at the school level are reported in parentheses.

*** Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

Table A.8
School-Specific Principal Experience

	$c =$ K5 vs. $c' =$ K8			$c =$ K5 vs. $c' =$ K6		
No. of years:	> 1	1	Δ	> 1	1	Δ
$\Phi_{c-c',5,post-pre}$	1.40**	1.07*	0.32	1.64**	1.75*	-0.11
	(0.67)	(0.65)	(0.92)	(0.79)	(0.93)	(1.08)
$\Phi_{c-c',5-4,post-pre}$	2.55***	1.16	1.40	1.37	0.54	0.83
	(0.81)	(0.78)	(1.09)	(0.92)	(1.05)	(1.38)

Note: This table presents difference-in-differences and triple-differences estimates for schools with principals who have two or more years of school-specific experience (> 1) and those who are new to the school (1), and the difference between the two (Δ). The estimates are constructed from joint F-tests of the interaction dummies (pre/post period × type × grade × established/new principal) included in the regression with full controls (specification (3)) for the subsample of schools that do not switch configuration during the period of analysis. Standard errors adjusted for clustering at the school level are reported in parentheses.

*** Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

Table A.9
Descriptive Statistics for 1998 Re-sorting Analysis

Variable:	Mean	St. Dev.	Min	Max
Transitions				
All Teachers	0.079	0.270	0	1
FE97: Above Median	0.066	0.248	0	1
FE97: Below Median	0.081	0.273	0	1
Gr. Scale Difference (Δg)	0.00	0.36	-2	2
Quality Measure (1997 FE)				
Above Median	0.52	0.50	0	1
Below Median	0.48	0.50	0	1

Note: Observations are at the teacher level and include only those teaching at K-5 schools that maintain their grade configuration from 1997 to 1998. The 1997 fixed effect quality measure is constructed using contemporaneous and prior scores, as well as demographic data.

Table A.10
Teacher Sorting by Quality (1998)

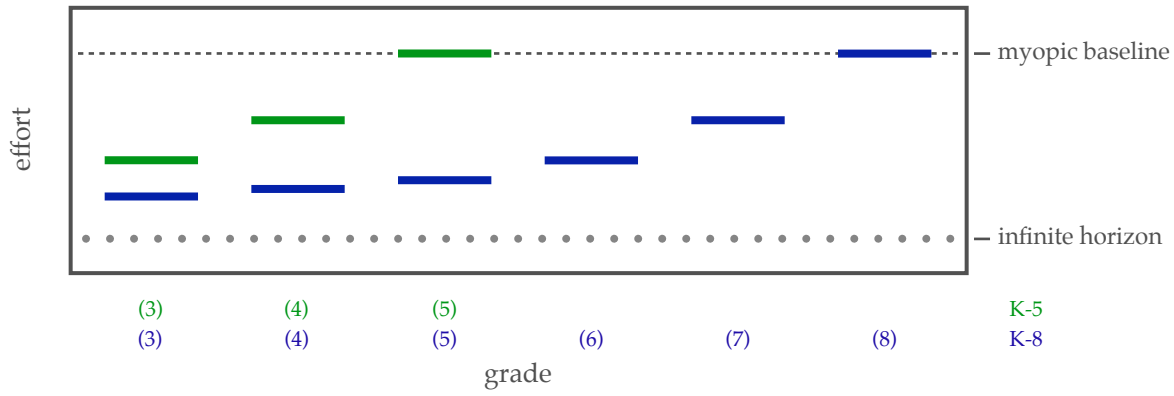
Gr. Scale Difference:	Δg_{97}^{h} (all tch)	Δg_{97}^{l} (all tch)	Δg_{97}^{h-l} (all tch)	Δg_{97}^{h-l} ($\Delta g \neq 0$)
Years of Experience x				
$1 \leq x \leq 12$	0.019	-0.021	0.040**	0.27**
	(0.013)	(0.013)	(0.019)	(0.12)
all x	0.006	-0.004	0.010	0.08
	(0.007)	(0.008)	(0.010)	(0.08)
$1 \leq x \leq 3$	0.051*	-0.034	0.085**	0.48**
	(0.029)	(0.025)	(0.038)	(0.21)
$4 \leq x \leq 6$	-0.008	-0.032	0.025	0.14
	(0.020)	(0.021)	(0.029)	(0.21)
$7 \leq x \leq 9$	0.018	-0.043*	0.061*	0.45*
	(0.024)	(0.025)	(0.035)	(0.26)
$10 \leq x \leq 12$	0.010	0.060*	-0.049	-0.30
	(0.031)	(0.034)	(0.045)	(0.34)

Note: Observations are at the teacher level and include only those teaching at K-5 schools that maintain their grade configuration from 1997 to 1998. Teacher quality measures are constructed from teacher fixed effects in a regression of test scores in 1997 on the prior ones in 1996 and set of covariates, including parental education, student ethnicity and student exceptionality. A high quality teacher (h) is then defined as possessing a fixed effect that is above the median, while a low quality teacher (l) possesses one that is below the median. The difference in grade scale (Δg) measures the change in grade and allows for changes in classes taught even if the teacher teaches multiple classes across different grades. Standard errors are reported in parentheses.

** Significant at the 5 percent level.

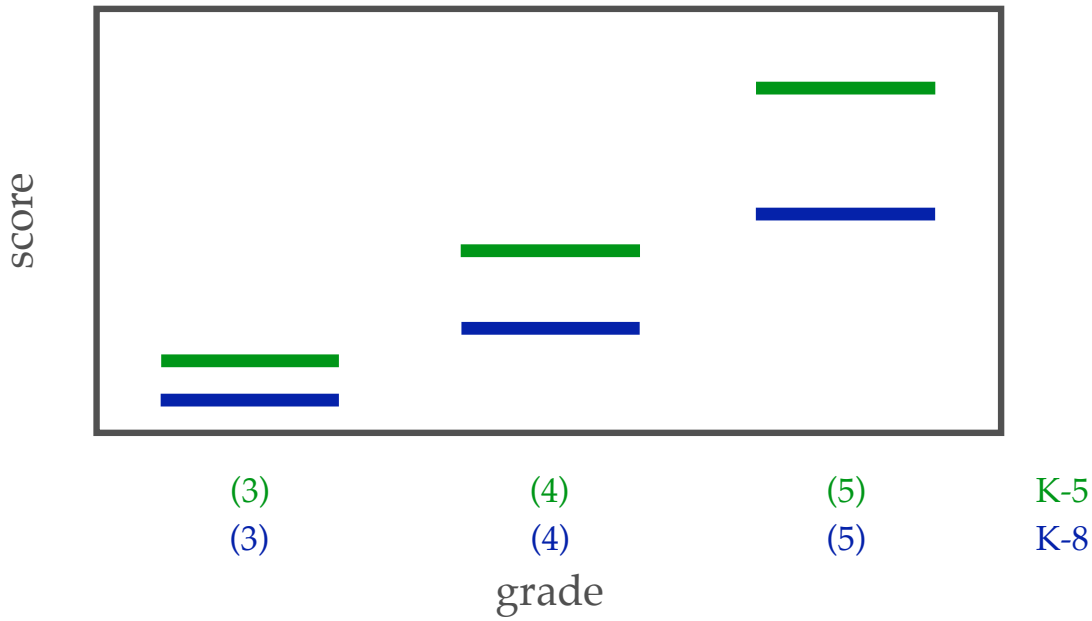
* Significant at the 10 percent level.

Figure A.1
A Comparison of Effort Between K-5 and K-8 Schools



Assuming that the target coefficient exceeds the natural growth rate ($\alpha > \gamma$), this diagram contrasts the effort levels by grade for two different grade spans (as implied by the first-order conditions). This is done to illustrate how differing horizons affect the effort level for a particular grade. In the final period, there is no future horizon to take into consideration. Thus, the effort level coincides with what would be chosen if agents were fully myopic. As the number of future grades increase, the effort response diminishes. In the limit, it is attenuated to the infinite horizon level of Weitzman (1980).

Figure A.2
A Comparison of Scores Between K-5 and K-8 Schools



Given the effort disparities predicted when $\alpha > \gamma$, this diagram provides an example of what the scores might look like by grade for two different grade spans that are identical in inputs. When comparing the scores across grade spans, two features should be evident. First, the score disparity is positive in favor of the school with the shorter horizon (K-5); second, the score disparity is increasing in the grade. These patterns are implied, respectively, by Propositions 1 and 2.

Figure A.3
Density of First-Differenced Scores By Grade (With Controls)

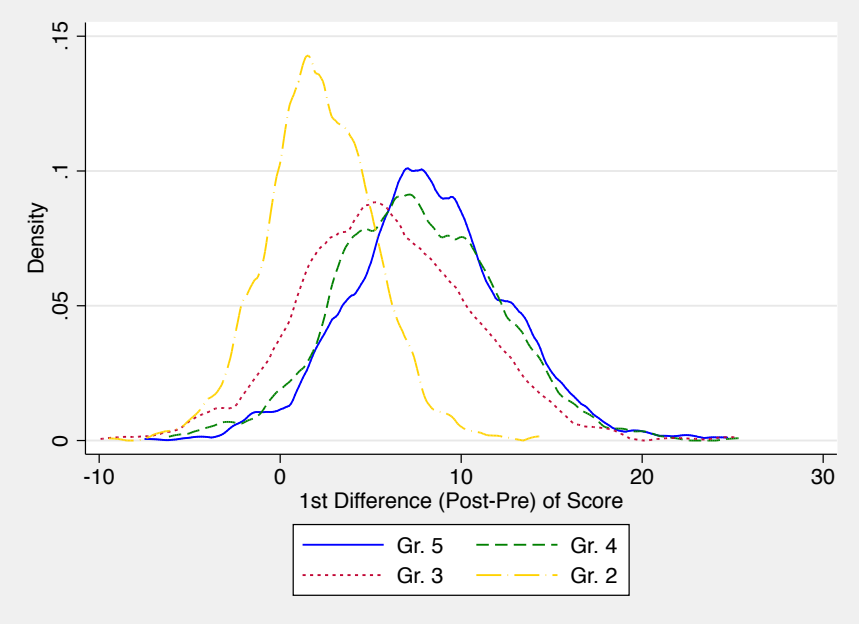


Figure A.4
Density of First-Differenced Scores By Grade (Without Controls)

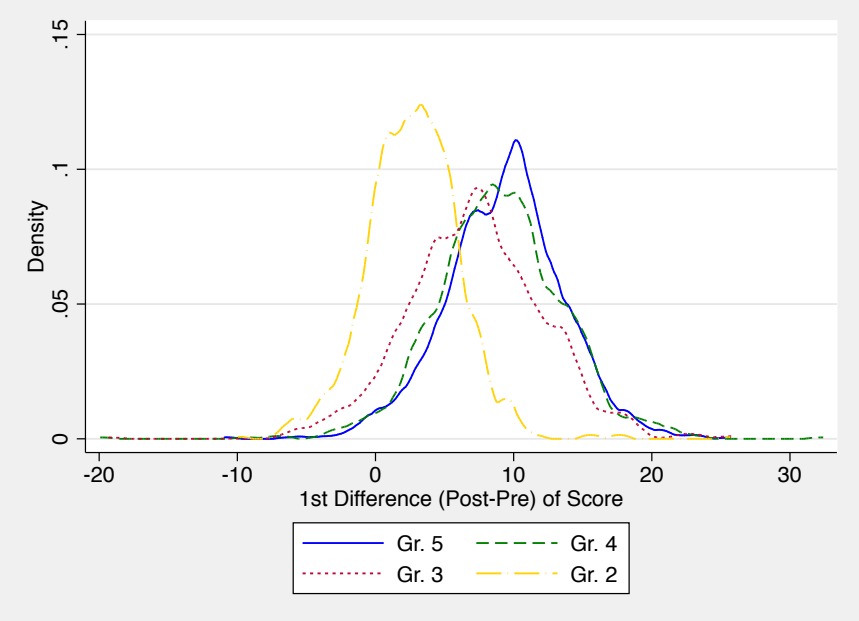
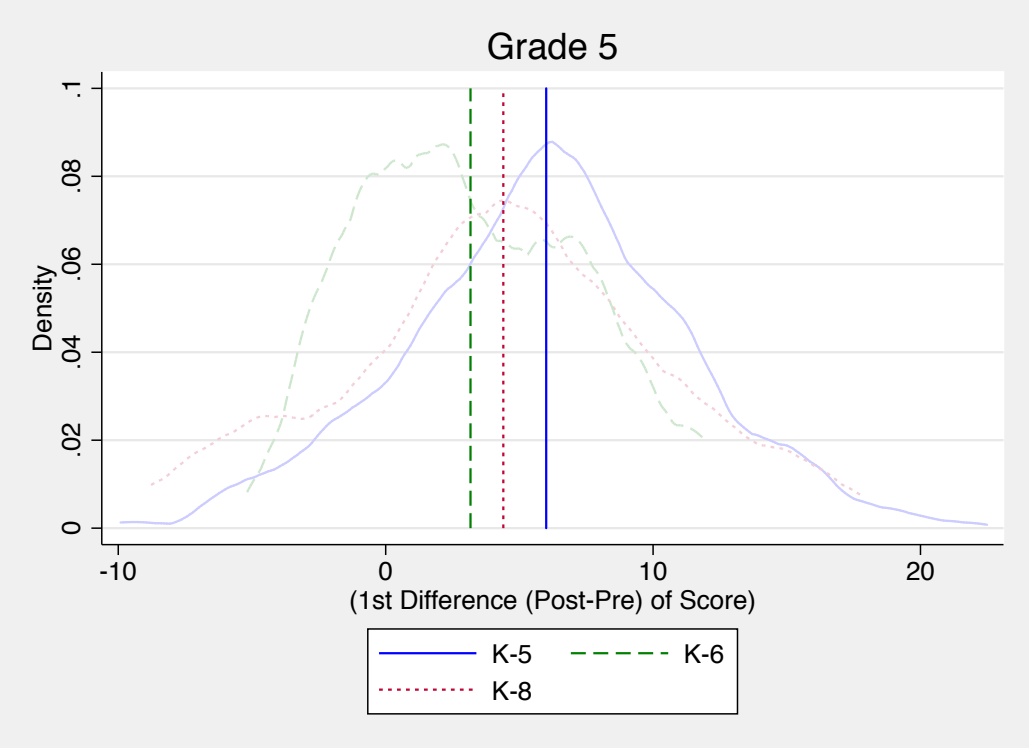


Figure A.5
Grade Five Distribution of First-Differenced Scores By Grade Span



This figure plots the density and means (given by the vertical lines) of the first-differenced grade five score by school configuration. As predicted by Proposition 1, K-5 schools have a higher mean than K-6 or K-8 schools. While it seems that the gain is lowest for K-6 schools, the K-6 and K-8 means are not statistically different from each other, since fewer K-6 observations lead to reduced power. In addition, the distribution of gains reflects selection bias arising from K-6 schools disproportionately switching to a new configuration, which I explicitly control for in Table 3.