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 and find 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, policy makers have increasingly embraced reforms that hold educators more

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. Contact the author at hugh.macartney@duke.edu. Information concerning access to the data used in this article is available as supplementary material online. Appendices A–F are available online.
accountable for the academic performance of their students. Such accountability measures have included conducting 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 nonmarginal 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 other states.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—this is 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

---

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.
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 all factors beyond a worker’s control so as to focus on pure effort. However, practical issues, such as ensuring that agents understand the scheme enough to respond to it and incomplete information, are likely to force the policy makers to select an imperfect target. In particular, they might intentionally neglect some available prior information for transparency reasons, or they may 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 that chooses effort in light of the prevailing incentives. The model 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

---

4 For example, they might only condition on the prior output of one agent, rather than all previous measures for all relevant agents.
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 5 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 school through high performance in grade 5 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 5 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 (e.g., grade 5 vs. grade 4) 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% of a standard deviation in the grade 5 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% of a standard deviation in the grade 5 score. To place these significant effects into context, the literature suggests that teachers account for between 8% and 15% of a standard deviation in test scores (see Rivkin, Hanushek, and Kain 2005; Rothstein 2010; Chetty, Friedman, and Rockoff 2014a). 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 through 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 that 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 that are determined by multiple agents in the system (e.g., bonus receipt for heterogeneous workers competing against each other in a tournament). My estimates demonstrate that there is a clear trade-off 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 policy makers should be cognizant of when designing incentive schemes.7

The remainder 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, Section V reports the main results, and Section VI carries out a set of robustness checks that address the key threats to validity. Section VII then explores likely mechanisms underlying the estimated dynamic effects, and Section VIII 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 reallocation, 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. (2014b). 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 in a realistic setting, I extend the dynamic moral hazard literature in this section. In particular, I focus on the strand that 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, as well as several testable predictions for the empirical investigation that follows.

A. The Environment

1. 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

---

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 (appendices A–F are 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 for additional detail on the reform.
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 $\alpha$ 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 intraschool incentives in the model.\textsuperscript{12}

Suppose there are $S$ schools, indexed by $s \in \{1, \ldots, S\}$, and let the grade within a particular school be referenced by $g \in G_c = \{0, \ldots, G_c\}$, 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.\textsuperscript{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 $\{e_{g,s}\}_{g \in G_c}$. Each choice $e_{g,s}$ is an input in the production of educational achievement for students and is selected from $\mathbb{R}^+$ according to the school’s payoff.

### 2. 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.\textsuperscript{14} At the end of every year $t$, a test is taken in this subject by all students in school $s$, generating average test scores for each school-grade pair. These scores are denoted by $y_{s,g}$ and are taken to be a measure of educational output for the relevant group of students and the representative teacher for that grade.

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 that students begin grade $g$ with using the prior score, $y_{s,g-1}$.\textsuperscript{12}

\textsuperscript{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.

\textsuperscript{13} For example, according to this notation, given that the receipt of the bonus in North Carolina depends on the scores for grades 3–8, $g = 0$ corresponds to grade 2 and $g = G_c = 3$ corresponds to grade 5 for a K–5 school (while I do not focus on earlier low-stakes grades here, grade 1 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 5 at a different school. Thus, $g \in G_c = \{1, 2, 3\}$ for such middle schools.

\textsuperscript{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.
Given this definition, I model the score $y_{s cg}$ as depending on the effort $e_{s cg}$ exerted by the representative teacher for the school-grade pair, the ability of the teacher $a_{s cg}$, the prior end-of-grade educational capital for current grade $g$ students $y_{s cg-1 t-1}$, and a grade-school-year shock $u_{s cg}$. Teacher effort and shocks are treated as common to all students within a classroom—this is 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

$$y_{s cg} = H(y_{s cg-1 t-1}, e_{s cg}, a_{s cg}) + u_{s cg},$$

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

$$y_{s cg} = \gamma y_{s cg-1 t-1} + e_{s cg} + a_{s cg} + u_{s cg}.$$  \hspace{1cm} (1)

3. 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 $y_{s cg}$ and targets $\hat{y}_{s cg} = \alpha y_{s cg-1 t-1}$ for each grade within the school, 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 intertemporal depreciation rate $\delta$, 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

$$U_{s ct} = \sum_{t=1}^{\infty} \delta^{t-1} \left\{ b \left[ 1 - F_{s c t} \left( \sum_{g=1}^{G_c} \left[ (\alpha - \gamma) y_{s cg-1 t-1} - e_{s cg} - a_{s cg} \right] \right) \right] - \sum_{g=1}^{G_c} C(e_{s cg}) \right\},$$

\hspace{1cm} (2)

\hspace{1cm} 15 Note that the multiplicative coefficient $\alpha$ is positive in my empirical application, as it is derived by regressing a current positive test score on a smaller positive prior one.
where $F_c(\cdot)$ is the cumulative distribution function of the sum of grade-specific shocks ($\sum_{g=1}^{G_c} u_{sgt} \equiv G_c \bar{u}_{st}$), and the benefit portion of the payoff function arises from the probability of receiving the bonus $\Pr[\sum_{g=1}^{G_c} y_{sgt} > \sum_{g=1}^{G_c} \bar{y}_{sgt}]$, which (using eq. [1]) is equivalent to $\Pr[\sum_{g=1}^{G_c} u_{sgt} > \sum_{g=1}^{G_c} ((\alpha - \gamma) y_{sgt - 1 - 1 - \epsilon sgt} - e_{sgt} - a_{sgt})]$.  

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 $\{e_{sgt}\}_{t=1}^\infty$ to maximize the objective in equation (2). Defining $\Pi_{st} = -\sum_{g=1}^{G_c} (e_{sgt} + a_{sgt} + (\gamma - \alpha) y_{sgt - 1 - 1})$ and assuming a quadratic cost function $C(e) = (d/2)e^2$, the first-order conditions that govern these choices are given by

$$
\frac{d}{de_{sgt}} = \begin{cases} 
    f_s(\Pi_{st}) + \delta(\gamma - \alpha) \sum_{j=0}^{G_c - g - 1} \delta^j \gamma^j f_s(\Pi_{st+1 + j}) & \text{for } 1 \leq g < G_c, \\
    \frac{d}{f_s(\Pi_{st})} & \text{for } g = G_c,
\end{cases}
$$

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, $f_s(\cdot)$ is the probability density function of $\sum_{g=1}^{G_c} u_{sgt}$, and $f_s(\Pi_{st})$ represents the school-specific contemporaneous incentives for period $t$ unrelated to the ratchet effect. 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 target coefficient exceeds the growth rate of the score ($\alpha > \gamma$).  

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

The proof is contained in appendix C.1 (appendices A–F are 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 ($\alpha < \gamma$).  

---

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.  
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.
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 $P$ terms drop away, leaving only ratchet effects that differ according to the school configuration and leading to expressions that are analytically tractable.\footnote{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).} The conditions become

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

\textsc{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 ($y_{cg} > y_{c'g}$, $\forall \ g \in G_c$).

The proof is contained in appendix C.2. 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, respectively, serve $G_c = 3$ and $G_c' = 6$ grades with high-powered incentives attached. As shown in the proof, the first-order 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 3, 4, or 5. Figure A.1 (figs. A.1–A.5 are 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. An analogous result holds for a comparison between K–5 and K–6 schools.

\textsc{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 $g$, $\forall \ g \in G_c$.\footnote{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).}
The proof is contained in appendix C.3. 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 5—this result is reflected in figure A.2 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).20 This includes information on North Carolina students, teachers, and schools for the years 1994–2005.21 Given that the accountability reform took effect in 1997, I refer to 1994–96 as pre-reform years and 1997–2005 as post-reform years. The data set contains yearly standardized test scores for each student in mathematics and reading from grades 2–8.22 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 3–8.

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

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 \( \alpha \) toward the growth rate \( \gamma \). Indeed, they are eliminated altogether if the planner sets \( \alpha = \gamma \). Intuitively, such a target coefficient no longer punishes teachers in the future for exerting higher effort today. However, there is a trade-off 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 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.

22 What are referred to as “grade 2” tests are administered in September of the grade 3 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.
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.

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 presents graphical evidence to show that test scores increase from the pre-reform period to the 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 5 (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.

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.
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 nonrandomly represented in the current population of schools.26 As a result, there is ample 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 5 score between K–5 and K–8 schools, for example, I would simply construct the difference-in-differences score

\[ \Delta \Delta y_{K5, K8, post-pre} = (y_{K5, post} - y_{K5, pre}) - (y_{K8, post} - y_{K8, pre}). \]

Such an approach adjusts for both preexisting 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

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 policy makers to shift grade 6 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.
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 that 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

$$y_{sgt} = X_{sgt}' \beta + \sum_{c=1}^{C} \sum_{g \in G} (\phi_{c,g,pre} + \phi_{c,g,post}) + e_{sgt},$$

where each $\phi$ is an interacted indicator variable that adjusts the score for every combination of grade, school type, and period. In essence, each fixed effect is a score for a particular school configuration and grade in the pre-reform or post-reform period, adjusted for the vector of observable controls.

Upon estimating equation (3), I use $F$-tests of the relevant $\phi$ 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

$$\Phi_{K5–K8,g,post−pre} = (\phi_{K5,g,post} - \phi_{K5,g,pre}) - (\phi_{K8,g,post} - \phi_{K8,g,pre}).$$

If unobserved trends are common across grade configurations, then $\Phi_{K5–K8,g,post−pre} > 0$ 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, as well as supply-side changes in the distribution of school configurations over time. One approach for

27 The results do not appreciably change when allowing for control coefficients to vary by grade ($\beta_{g}$) 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 Sec. III, North Carolina policy makers 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 indicate that switching K–8 schools have a lower score than those that do not switch, the direction of such bias is likely downward.
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 that 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 comparing grade 4 and grade 5 scores between K–5 and K–8 schools is

\[
\Phi_{K5-K8.5; 4, post-pre} = \Phi_{K5-K8.5; post-pre} - \Phi_{K5-K8.4; post-pre}
\]

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

Such an analysis controls for time-invariant effects and shared trends between configurations, but it 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-side or supply-side selection bias is addressed by the triple-differences approach. A finding of \(\Phi_{K5-K8.5-4, post-pre} > 0\) 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 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 \(X_{st}\). These specifications are defined in table A.1 (tables A.1–A.10 are 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).

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.
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 1–3 in table A.1 and grades 3–5, 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 for K–5 and K–6 schools are reported in table 2. In every case, the difference between pre-reform and post-reform scores for a specific configuration is positive and significant, consistent with the descriptive evidence. Using specification 3, the pre-to-post gains in grade 5 scores for K–5, K–8, and K–6 schools are 8.8, 7.3, and 5.9 developmental scale points, respectively. The analogous gains in grade 4 scores are 7.5, 7.1, and 5.6 points and in grade 3 scores, 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 5 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 5 to grade 4 and positive but smaller for the comparison between grades 4 and 3 when including controls. This pattern across grades is precisely what is predicted by the theory (proposition 2).

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 grades 3 and 4 than the 16.9 developmental points in grade 5 (see table 1), this pattern becomes even more pronounced when adjusting for variation in scores.
33 The estimates for grades 3 and 4 are also positive, but not significantly so.
Table 2
Main Results

<table>
<thead>
<tr>
<th></th>
<th>$c = K-5$ versus $c' = K-8$</th>
<th>$c = K-5$ versus $c' = K-6$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Grade 5 difference-in-differences:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Phi_{t,5,post-pre}$</td>
<td>9.01*** 9.51*** 8.77*** 9.01*** 9.51*** 8.77***</td>
<td>(.22) (.15) (.14) (.22) (.15) (.14)</td>
</tr>
<tr>
<td></td>
<td>(8.41*** 7.98*** 7.31*** 7.30*** 6.76*** 5.92***</td>
<td>(.36) (.29) (.31) (.45) (.29) (.32)</td>
</tr>
<tr>
<td>$\Phi_{t'=5,post-pre}$</td>
<td>.60 1.53*** 1.46*** 1.71*** 2.75*** 2.84***</td>
<td>(.43) (.33) (.34) (.52) (.33) (.35)</td>
</tr>
<tr>
<td>Grade 4 difference-in-differences:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Phi_{t,4,post-pre}$</td>
<td>7.76*** 8.10*** 7.52*** 7.76*** 8.10*** 7.52***</td>
<td>(.19) (.13) (.14) (.19) (.13) (.14)</td>
</tr>
<tr>
<td></td>
<td>(7.96*** 7.50*** 7.05*** 6.29*** 6.02*** 5.62***</td>
<td>(.43) (.36) (.42) (.51) (.33) (.37)</td>
</tr>
<tr>
<td>$\Phi_{t'=4,post-pre}$</td>
<td>-.20 .60 .47 1.47*** 2.08*** 1.89***</td>
<td>(.47) (.38) (.45) (.56) (.36) (.40)</td>
</tr>
<tr>
<td>Grade 3 difference-in-differences:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Phi_{t,3,post-pre}$</td>
<td>7.21*** 7.48*** 6.79*** 7.21*** 7.48*** 6.79***</td>
<td>(.20) (.15) (.16) (.20) (.15) (.16)</td>
</tr>
<tr>
<td></td>
<td>(7.26*** 7.27*** 6.74*** 5.96*** 5.92*** 4.94***</td>
<td>(.46) (.40) (.41) (.48) (.33) (.38)</td>
</tr>
<tr>
<td>$\Phi_{t'=3,post-pre}$</td>
<td>-.05 .21 .04 1.25** 1.55*** 1.85***</td>
<td>(.51) (.43) (.45) (.53) (.36) (.41)</td>
</tr>
<tr>
<td>Triple-differences:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Phi_{t'=c-3,post-pre}$</td>
<td>.80*** .93*** .99*** .24 .66*** .95***</td>
<td>(.38) (.34) (.44) (.31) (.27) (.36)</td>
</tr>
<tr>
<td>$\Phi_{t'=c-4,post-pre}$</td>
<td>-.15 .39 .42 .22 .53 .04</td>
<td>(.38) (.41) (.47) (.33) (.34) (.44)</td>
</tr>
</tbody>
</table>

**NOTE.**—For each specification defined in table A.1 (and in words in the first paragraph of Sec. 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 5% level.
*** Significant at the 1% level.

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 5, depending on the control-based specification used. This is equivalent to an effect that is between 8.6% and 9.1% of a standard deviation in the grade 5 score. For the triple-differences estimates, the effect is estimated to be between 0.80 and 0.99 scale points for the grade 5 to grade 4 comparison (or between 4.7% and 5.9% of a standard deviation in the grade 5 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% of a standard deviation in the grade 5 score).  

VI. 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 5 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 5 to grade 4 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 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.

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 and show that the dynamic gaming effects are most pronounced for mathematics test scores in table A.6, 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, e.g., Rivkin et al. 2005).
The Dynamic Effects of Educational Accountability

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 specifications 1 and 3. In all cases, the estimates are not statistically different from each other. Given that specification 1 includes no controls while specification 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 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 (e.g., educators in rural areas may respond to incentives differently than their urban or suburban counterparts). Given the strength of the findings for the 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. 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 (vs. 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

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% 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% of teachers who change schools each year.
with conventional public schools in 1998, student accountability in 2001, 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 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 5 difference-in-differences and grade 5 to grade 4 triple-differences estimates are largest in the actual year of the reform for the K–5 and K–8 and the K–5 and K–6 comparisons. 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.

VII. 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. Understanding 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.

37 From Bifulco and Ladd (2006), the number of charter schools in 1998 was 27, growing to 67 by 2002.  
38 Fifth-grade students (and third-grade students, beginning in 2002) 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.  
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.
In appendix F.3, 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

Table 4
Decomposing the Dynamic Effects

<table>
<thead>
<tr>
<th></th>
<th>Unadjusted for Share</th>
<th>Adjusted for Share</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Phi_{k \cdot \text{post-pre}}$</td>
<td>$\Phi_{k \cdot \text{post-pre}}$</td>
</tr>
<tr>
<td>A. Estimates for K–5 and K–8 comparison:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta g = 0$</td>
<td>1.14*** 1.34*</td>
<td>.60** .71*</td>
</tr>
<tr>
<td></td>
<td>(.53) (.71)</td>
<td>(.28) (.37)</td>
</tr>
<tr>
<td>$\Delta g \neq 0$</td>
<td>1.13* 1.49*</td>
<td>.53* .70*</td>
</tr>
<tr>
<td></td>
<td>(.63) (.87)</td>
<td>(.30) (.41)</td>
</tr>
<tr>
<td>Total effect</td>
<td>1.13*** 1.41**</td>
<td>(.46) (.55)</td>
</tr>
</tbody>
</table>

B. Estimates for K–5 and K–6 comparison:

<table>
<thead>
<tr>
<th></th>
<th>Unadjusted for Share</th>
<th>Adjusted for Share</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Phi_{k \cdot \text{post-pre}}$</td>
<td>$\Phi_{k \cdot \text{post-pre}}$</td>
</tr>
<tr>
<td>$\Delta g = 0$</td>
<td>1.42*** 1.05</td>
<td>.76** .56</td>
</tr>
<tr>
<td></td>
<td>(.67) (.74)</td>
<td>(.36) (.39)</td>
</tr>
<tr>
<td>$\Delta g \neq 0$</td>
<td>1.94*** .96</td>
<td>.91** .45</td>
</tr>
<tr>
<td></td>
<td>(.78) (.76)</td>
<td>(.36) (.35)</td>
</tr>
<tr>
<td>Total effect</td>
<td>1.67*** 1.01</td>
<td>(.61) (.62)</td>
</tr>
</tbody>
</table>

**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 $\times$ type $\times$ grade $\times$ 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-reform and post-reform periods are, respectively, defined as 1995–96 and 1998–2005. 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 10% level.
** Significant at the 5% level.
*** Significant at the 1% level.

In appendix F.3, 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

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.
(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 $\Delta g = 0$.

Under the more restrictive pre-reform 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.

**VIII. 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% of a standard deviation in the grade 5 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 policy makers 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


