The Dynamic Effects of Educational Accountability *

Hugh Macartney †
Department of Economics
Duke University

October 16, 2012

Abstract

Recent education accountability reforms feature school-level performance targets that condition on prior scores to account for student heterogeneity. Yet doing so introduces potential dynamic distortions to incentives: teachers may be less responsive to the reform today to avoid more onerous future targets — an instance of the so-called 'ratchet effect.' Guided by a dynamic model and utilizing rich educational panel data from North Carolina, I exploit school grade span variation to identify any dynamic gaming, finding compelling evidence of ratchet effects. I then directly estimate the structural parameters of the corresponding model, uncovering complementarities between teacher effort and student ability.

Keywords: Public, Education, Personnel, Dynamic Gaming, Dynamic Incentives, Ratchet Effects, Education Production, Educational Accountability

JEL Classifications: D82, I21, J24, J33, M52

* I would like to thank Robert McMillan, Aloysius Siow and Carlos Serrano for their guidance and support throughout this project. Thanks also to Douglas Almond, Michael Baker, Dwayne Benjamin, Sandra Black, Gustavo Bobonis, Branko Boskovic, Raj Chetty, Damon Clark, Stephen Coate, Elizabeth Dhuey, Weili Ding, Raquel Fernández, Amy Finkelstein, Christopher Flinn, Sacha Kapoor, Steven Lehrer, Thomas Lemieux, Joshua Lewis, Enrico Moretti, Alvin Murphy, Parag Pathak, Nancy Qian, Petra Todd, Trevor Tombe, Jacob Vigdor, and participants at the 4th Annual CLSRN Conference in Quebec City and the CEPA seminar at the University of Toronto 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.

† Email: hugh.macartney@duke.edu

1 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 introducing 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 these targets are constructed is of particular interest from an incentive design perspective. Simple schemes, such as the one used under the federal No Child Left Behind Act of 2001, set performance targets that are independent of student, teacher, or school measures — past or present. In contrast, more refined value-added schemes feature targets that condition on prior scores to adjust for input heterogeneity. For instance, under North Carolina's sophisticated accountability system, established in 1996, all teachers and the principal at a school receive a monetary bonus if the school meets specified growth targets in student achievement, these targets conditioning on student prior test scores.[1]

Despite the clear benefits of the value-added approach, targets that depend on lagged achievement are potentially manipulable over time. In particular, raising effort under a scheme such as North Carolina's not only affects the likelihood of exceeding the current target, but also determines the target that follows, so that a strong performance today makes it more difficult to reap a bonus tomorrow. Given this knowledge, teachers 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.'

The central goal of this paper is to measure the extent to which such dynamic distortions matter in practice. As a starting point, it is instructive to turn to the substantial theoretical literature that explores dynamic moral hazard issues. In the seminal paper by Weitzman (1980), workers make effort choices facing an infinite horizon, where targets depend on earlier output.[2] The main prediction to emerge is intuitive — that agents should identically suppress effort in every period. Yet the theory does not lend itself in a straightforward way to empirical testing, as this prediction is indistinguishable from static gaming period-by-period.

With a view to obtaining predictions relating to dynamic effects that can be assessed

[1] Another example is the 1999 California accountability reform, which conditioned targets on the prior scores associated with given teachers. It was discontinued shortly after its introduction due to a budget shortfall.

[2] See also Holmstrom (1982) and Keren *et al.* (1983) for an analysis of the ratchet effect under a fixed sub-optimal target without renegotiation. Freixas *et al.* (1985), Lazear (1986), Baron and Besanko (1987), Gibbons (1987), Laffont and Tirole (1988), and Kanemoto and Macleod (1992) address ratchet effects under various mechanisms with limited or no commitment.

empirically, I develop a theory modeled in a stylized way on the North Carolina reform. The theory features incentive targets that depend on the average prior score of students — in practice, the target given to each school aggregates grade-specific targets that are proportional to the average score of individual students in the prior grade and year. For ratchet effects to exist when student prior scores determine the target, teachers must collectively respond to the school-level incentives to some degree. While this may occur without overt coordination, the mechanism that I envision is one where each principal centrally coordinates and monitors teachers to maximize her school's payoff.[3] Thus the decision makers in the model are school principals, reflecting the fact that actual incentives are at the school level. In this setting, the relevant horizon for dynamic gaming is finite rather than infinite. This is because students only attend a particular school for a fixed period of time, and the contribution of a student to the school aggregate target persists only as long as the student remains in the school.[4]

The theory generates a crucial insight: the extent of gaming is predicted to vary according to the horizon faced by the school principal. Intuitively, when the horizon becomes shorter, the downside associated with outstanding performance is mitigated since there are fewer periods in which the target will be raised in future, and so teachers will tend to increase their effort. Moreover, in the limiting case with a horizon consisting of only a single period, there would be no future targets to consider, leading any dynamic distortions to disappear completely. Alternatively, I show that the ratchet effect can be eliminated in any multi-period setting in the special case where the target coefficient is identical to the natural growth rate of the underlying production process. If the next-period target can be met without any additional effort tomorrow, the incentive to dynamically game the system by distorting effort today is removed.

In the context of the North Carolina reform, the principal's horizon is captured well by the grade span of the school.[5] Given that I observe multiple grade-span configurations, this 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

[3]Principals may engage in the within-school re-assignment of teachers to classes according to teaching ability, in addition to influencing teacher effort. Such a possibility is considered in a companion paper of mine.

[4]Strictly speaking, it persists up until the penultimate grade the student is in the school.

[5]In North Carolina, students in kindergarten through grade eight are served primarily by one of three types of school structure. The majority of students first attend a K-5 school, which serves them through grade five, and then move to a 6-8 middle school. Others remain in elementary school until grade six at a K-6 school before progressing to a junior high school. In the third type, students attend the same school until grade eight, termed K-8 schools.

level of effort than their K-8 or K-6 counterparts, leading to a positive score differential in favor of K-5 schools.

The reasoning is as follows, building on the prior logic: In the case of a K-5 school, effort affects the probability of obtaining a reward today and also influences the grade six target that a separate 6-8 school faces tomorrow, since grade five is the final grade served by the K-5 school. Therefore, there will be no ratchet effect in grade five at the K-5 school. In contrast, a K-8 school serves grade six students as well, meaning that both the grade five and six outcomes matter for satisfying the overall target across all tested grades. Whereas the K-5 school imposes a negative externality on a 6-8 school, the K-8 school will internalize this externality by responding less to the scheme in the fifth grade to ensure a more attainable target in grade six.

To obtain evidence of distortions, a simple comparison of mean scores across different configurations could be misleading since the average school in each configuration may differ along other dimensions unrelated to the ratchet effect — K-5 schools might possess more able students and teachers than K-6 schools, for instance.[6] Given the possibility of unobserved differences between grade structures, I first employ a difference-in-differences estimation strategy, taking advantage of score data before and after the reform to identify the predicted dynamic gaming effect. Under the assumption that all differences in inputs and technology between two grade configurations are time-invariant, any change in the score disparity can be attributed to differential effort choices arising from the implementation of the scheme. All other disparities are removed through differencing. The initial descriptive evidence indicates that K-5 schools do indeed experience greater growth in grade five scores than either K-6 or K-8 schools once the reform is implemented.

To account for potential bias arising from differentially trending unobservables across school configurations, I then compute triple-differences estimates, exploiting score variation across grades as well as over time. This preferred analysis reveals substantial distortions between K-5 and K-8 schools — between 4.7% and 5.9% of a standard deviation in the grade five score in favor of the shorter grade span. The analogous distortion in grade five for the comparison between K-5 and K-6 schools is between 3.9% and 5.6% of a standard deviation. These findings are consistent with the predictions of the model and, importantly, are obtained without having to make overly restrictive identifying assumptions.

Using robustness checks, I am able to discount alternative explanations for the effects

[6]In addition, the production technology governing growth rates in scores may differ across configurations due to divergent peer effects, stemming from the presence or absence of older students in the school. See Cook *et al.* (2008) and Bedard and Do (2005) for a discussion.

I uncover. Notably, a falsification exercise rules out other education reforms that were implemented during the post-reform period as a driver of the results. Moreover, a subsample restriction that includes only those schools that maintained their configuration throughout the period of interest guards against selection bias arising from schools changing their grade span. In addition, an analysis of the estimates by subject lends credence to the interpretation that the dynamic effects are due to the accountability reform, rather than differentially trending peer effects by grade. Apart from identification concerns, I also consider coordination issues among multiple teachers and the role that the principal plays in responding to incentives.

Beyond using the model to obtain reduced-form evidence of dynamic target manipulation, the linkage between theory and data permits a more sophisticated analysis. In particular, an attractive feature of the approach is that key structural parameters of the model can be inferred directly from the robust difference-in-differences results, using a linear technology assumption. Accordingly, I obtain parameter estimates under a model with fully persistent educational inputs and also one where the teacher contribution to student learning is partially transitory. With those estimates in hand, illuminating counterfactual policy simulations can then be carried out directly, exploring the benefits of the existing scheme and the cumulative effects of ratcheting behavior.

Rather than following that course when considering counterfactuals, I adopt a more general approach. Using reduced-form estimates to infer the underlying structure of the model requires, as mentioned, a relatively strong linearity assumption to be made; and although that exercise is informative, it would be interesting to know whether nonlinearities are important in practice. Such a generalization is possible by virtue of the rich data at my disposal and the concrete predictions of the model. Allowing for a nonlinear interaction between teacher effort and student ability in production, the parameters are identified through variation in effort across the grade horizon within and across schools, and are estimated using a maximum-likelihood estimation approach. The nonlinear specification I choose also allows one to test between the linear and nonlinear technology variants. Upon estimating this more general model, I find evidence that effort and student ability, as proxied by the student prior score, are complements in the production of learning.

Taking the results of this analysis in combination with the model, I conduct two policy experiments. The first reveals the substantial effects of the reform, where the average cumulative grade five score in K-5 schools would be 44.3% of a standard deviation lower without the accountability scheme. Based on a key prediction of the model, the second experiment then explores a world in which the ratchet effects are eliminated entirely. Doing so results in

an average grade five score that is 1.7% of a standard deviation higher, but is also around 37% more costly to implement, owing to the fact that the theoretical prescription is to lower the target, which then makes it easier to satisfy. Further, a comparison of the counterfactual results under linear and nonlinear specifications reveals that the former overstates the cumulative effect of eliminating ratcheting behavior by 5.8%, thereby emphasizing the importance of the nonlinear generalization.

The rest of the paper is organized as follows: The next section reviews the relevant prior literature. Section 3 presents a simple theoretical model of dynamic gaming that yields the central insight used subsequently to estimate dynamic distortions. Section 4 discusses the 1996 North Carolina accountability reform in greater detail, and Section 5 describes the data, presenting stylized facts regarding the aggregate impact of the reform. Section 6 outlines the reduced-form econometric framework, reports the associated results and considers threats to their validity. Section 7 moves beyond such an analysis by deducing the structural parameters of the model directly from reduced-form estimates. Then in Section 8, I estimate a more general variant of the underlying production technology with nonlinearities in inputs, which yields evidence of complementarities in production. Section 9 describes the outcomes of the two counterfactual policy experiments, and Section 10 concludes.

2 Prior Literature

The current paper contributes to three main strands of research. The first is the dynamic moral hazard literature that analyzes the ratchet effect from a theoretical perspective. Weitzman (1980), Holmstrom (1982) and Keren *et al.* (1983) consider the ratchet effect when the planner commits to a suboptimal incentive scheme that features a revision procedure. Subsequent research, including that by Freixas *et al.* (1985), Lazear (1986), Baron and Besanko (1987), Gibbons (1987) and Laffont and Tirole (1988), has explored ratcheting behavior under mechanisms with limited or no commitment, while Kanemoto and Macleod (1992) consider ratchet effects in the presence of labour market competition. Motivated by the institutional details of the educational accountability reform in North Carolina, I build on this strand of literature by considering finite-period ratcheting behavior under a specified revision procedure. By focusing on the finite horizon, the theory yields a new insight into the identification of ratchet effects as well as several testable predictions for the empirical analysis to follow.

Building on existing theory, there is also a small empirical literature measuring ratchet effects. On the experimental side, Cooper *et al.* (1999) find evidence of a ratchet effect using

Chinese students and managers, while Charness *et al.* (2010) determine that embedding market competition for agents and principals in their experiment using undergraduate students decreases ratcheting behavior, which is in line with the prediction of Kanemoto and Macleod (1992) to the effect that increased competition attenuates the ratchet effect. With respect to observational evidence, Parent (1999) analyzes data from the National Longitudinal Survey of Youth and uncovers variation that is consistent with the ratchet effect. In particular, he exploits categorical data on the types of pay-for-performance used in the workplace for each respondent, if any, and finds that wages tend to be higher for piece-rate workers earlier in their career. This is in accordance with a prediction from Lazear (1986). In another study, Allen and Lueck (1999) detect some limited evidence of ratcheting behavior using a cross-sectional agricultural dataset. I contribute to this strand of literature by analyzing a specific large-scale incentive scheme using panel data and exploiting a novel source of identifying variation, associated with differences in the horizon faced by agents.

The third strand is a large literature on educational accountability, which can be further subdivided into three categories that are relevant to my work. The first category is concerned with evaluating accountability programs to determine if they have the desired effect on student achievement. Using cross-state variation in accountability strength, Carnoy and Loeb (2002) and Hanushek and Raymond (2005) find, independently, that test scores are higher under more accountable systems. Using the results of a survey that focuses specifically on pecuniary aspects of accountability, this finding is echoed by Figlio and Kenny (2007). As for assessing particular monetary reward schemes, Lavy (2002, 2009) utilizes data on Israeli schools to provide convincing evidence that performance-contingent bonuses lead to improved educational outcomes, while Muralidharan and Sundararaman (2009) conduct a large-scale randomized experiment in India and find that heightened incentives give rise to substantially higher test scores.[7] In one of my counterfactual policy experiments, I also provide evidence indicating that greater accountability has a positive effect on student performance.

The second category in the accountability literature concerns teachers gaming the system. Ladd and Zelli (2002) present the results of a survey suggesting that principals redirected resources from untested to tested subjects in response to greater accountability in North Carolina. Supplementing such survey evidence, a number of studies have detected gaming in test score data. Cullen and Reback (2006) assess the practice of exempting disadvantaged students from testing under the Texas accountability system, while Neal and Schanzenbach (2010) re-

[7]The authors rule out differential teacher attendance as a primary driver of their results, given that control and treatment schools are similar in this dimension. They reason that this leaves teacher effort as the most likely channel through which teachers respond to the scheme.

veal evidence consistent with Chicago teachers 'teaching to the distribution' of students. In addition, Jacob and Levitt (2003) demonstrate that overt cheating by teachers occurred in response to greater accountability in Chicago schools. My work builds on this literature, focusing on a form of gaming that occurs through a dynamic channel.

In very recent work, Barlevy and Neal (2011) propose an elegant theoretical method for dealing with many forms of gaming by eliminating the reliance of incentive schemes on cardinal-based measurements. In particular, they suggest using peer-to-peer contests between comparable students to form ordinal rankings of performance across teachers. They show that basing teacher compensation on such rankings results in efficient levels of effort by teachers. Given that relative performance is all that matters under the system they propose, tests with completely new content can be administered each year, thwarting undesirable 'teaching to the test.'

The third category in the accountability literature seeks to understand the mechanisms behind successful programs and, in doing so, determining whether and how they can be improved upon. Several studies are concerned with the basic methodology underlying the inference of teacher effects from score data. Considering numerous alternative specifications, some of which form the basis for existing high-powered incentive targets, analyses such as Todd and Wolpin (2003, 2007), McCaffrey *et al.* (2004) and Rothstein (2010) conclude that strong assumptions are needed in order to identify teacher effects, noting that bias in such estimates may arise for a variety of reasons.[8] In a specific experimental setting, Kane and Staiger (2008) show that the bias arising from non-random matching between teachers and students may not be as high as predicted by non-experimental analyses. In particular, the authors cannot reject the hypothesis that value-added estimates from pre-experiment data are unbiased measures of the true teacher value added under randomized classroom assignment. Addressing a different source of bias, Kane and Staiger (2001) propose an incentive scheme to filter out unwanted transitory processes, such as period-specific shocks arising from sampling variation, by averaging over multiple prior periods of performance and adjusting for differences in class and school size. Ahn (2009) employs a more direct way to infer teacher effects, harnessing variation in teacher absences and student test scores to infer teacher effort, also making use of the North Carolina accountability data. Under the intuitive and plausible hypothesis that teachers exert greater effort when their actions matter more at the margin for receiving a bonus, he finds that absences — assumed to vary inversely with teacher effort — are fewer when the difference

[8]For instance, bias will arise under a value-added specification if the grade-to-grade decline in an educational input's effect on the score in not the same across all inputs; it will also arise if assignment of teachers to classes varies non-randomly with other predictors of learning.

between the score and target (his proxy for incentive strength) is small and greater when the difference is large.[9]

I contribute to this last category of literature in several ways. First, I develop a detailed model that embeds many of the institutional details of the North Carolina reform, including the potential manipulability of targets. By structurally estimating the model, I uncover valuable information about the underlying learning technology, finding that nonlinearities matter in production. I also gain a better understanding of the assumptions required for identifying teacher effects in a dynamic setting, such as imposing restrictions on the growth and interaction of scholastic inputs in the evolution of student learning. Lastly, I explore the scope for improving the existing policy by proposing and implementing an alternative scheme that eliminates ratcheting behavior.

3 Theoretical Model

There are several reasons for extending the theoretical dynamic moral hazard literature. First, doing so allows me to develop intuition as to the possible workings of the ratchet effect in a setting where the horizon is finite and of varying length. This is in contrast to the infinite- and two-period models considered in the bulk of the pre-existing literature. By emphasizing the finite horizon, the theory yields a new insight concerning the identification of ratchet effects in such a setting, in addition to several testable predictions for the reduced-form investigation that follows. Moreover, since there is a mapping between the model and data by design, much more can be done. In particular, the model's structural parameters can be recovered directly from the reduced-form estimates, using a linear technology assumption. Knowing the parameters is valuable as it permits a more sophisticated analysis, in which counterfactual policy experiments can be carried out. With respect to such experiments, the model is once again informative in that it provides a specific recipe for refining the scheme to eliminate ratcheting behavior. Moving beyond the simple linear technology assumption, nonlinearities in the production of learning can also then be explored, exploiting the key structure of the theoretical model and estimating an econometric variant of it directly.

In this section, I present a simple theoretical framework that applies to a stylized education context, modeled on the North Carolina case as described in Section 4 below. In that

[9]Given that targets under the North Carolina accountability reform depend on student prior scores, there is almost certainly a correlation between the contemporaneous score and the target, making the incentive strength measure in Ahn's study potentially endogenous. The current paper focuses on the manipulability of this target.

setting, school-level incentive targets depend on student prior scores, and a single agent (the school principal) coordinates actions across grades, generating differential behavior according to the school's grade horizon.[10] The model is related to the seminal paper by Weitzman (1980), which predicts the emergence of ratchet effects when performance today determines bonus receipt today and tomorrow.[11] 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,' not 'school principal,' sense) must reward agents, conditional on current and prior output. To see this, consider an agent's problem at time t. Given the scheme and a convex cost of output $C(\cdot)$, this 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 $\delta\alpha > 0$.[12] 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).

While the basic idea of Weitzman (1980) is contained in my model, my formulation differs in several respects. As noted above, I consider ratchet effects in a finite-period setting, reflecting the fact that school-level targets depend on student prior scores in North Carolina and that students do not attend the same school forever. Another important difference is that, in addition to the contemporaneous choice of teacher effort, output 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.[13] This means that even

[10] More generally, the model is easily adapted to any environment where an agent faces a value-added scheme and is only responsible for output over a finite number of periods.

[11] 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.

[12] 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 in this paper, for example).

[13] 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

if the target does not depend on the prior score, the current choice will still affect all future output levels. In addition, I allow for the possibility that incentives are nonlinear, which is suggested by the type of threshold-based incentives employed in practice. I now describe this model in greater detail and use it to develop testable predictions.

3.1 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 possess the means of calculating the school-level target, which is relatively straightforward 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.[14] Let there be S schools, indexed by $s \in \{1, \ldots, S\}$, and let each 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 school s with grade configuration c, normalized so that $g = 1$ is the first grade with high-powered incentives attached.[15]

In any given year t, each school s with a finite horizon dictated by its configuration c chooses a set of effort levels $\{e_{scgt}\}_{g \in \mathcal{G}_c}$, which are inputs in the production of educational achievement for students. Each choice e_{scgt} is selected from the set of continuous effort levels according to the school's preference ordering over them. This ordering is determined by the production function, which converts a particular level of effort into educational output, the incentive scheme that is selected by the planner, and the cost of effort.

time, appropriate given the cumulative nature of the education process.

[14] This modeling choice is made to focus on the core idea of ratcheting behavior. It assumes that the principal is capable of perfectly co-ordinating her teachers. If this were not the case, then ratchet effects would be attenuated. However, since ratcheting is very apparent in my empirical analysis, free-riding cannot be significantly impeding co-ordination, suggesting that this assumption is reasonable.

[15] For example, consider a K-5 school, which serves students in kindergarten through grade five. 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 for a K-5 school (the grade prior to high-powered incentives being introduced), while $g = 3$ corresponds to grade five, the last grade served. For a 6-8 school, there is no grade for which $g = 0$, as it represents grade five at a different school. Thus, in this case, $g \in \{1, 2, 3\}$.

Inputs and Production Technology

For simplicity, I abstract away from the two tested subjects used in practice in North Carolina by assuming that there is a single representative subject.[16] 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.

The education process 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 in learning, such as the student's raw intelligence and the contributions of her teachers, have a lasting impact on her ability to learn in the future. As these prior inputs are not directly observed, I summarize the prior end-of-grade educational capital with which students begin grade g using the prior score $y_{scg-1t-1}$.[17]

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}. In the model, teacher effort and shocks are treated as common to all students within a classroom, which is a reasonable assumption to make 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. Initially, I also consider the effect of teacher effort on student development to be permanent so that it affects the subsequent score in the same way as educational capital.[18] In general, let the student's score in school s, grade configuration c, grade g and time t be given by

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

[16] This assumption can be made without loss of generality, since any dynamic effects that arise should be manifested in both scores, given that a bonus is only awarded if the school-level composite target is satisfied. The modeled one-subject test score can be conveniently interpreted as the sum of the reading and mathematics scores.

[17] In practice, this will be a noisy measure of educational capital. However, given that the empirical analysis is at the school configuration level, this will only bias results if the expectation of such noise differs across grade structures.

[18] Later, I consider the implications of allowing effort to be partially transitory, which would occur if teachers choose to devote some of their effort to 'teach to the test,' for instance.

which potentially allows for teacher effort and the capital stock of the average student to interact in the production of learning. Although such an interaction may be a more realistic representation of educational production for the purposes of predicting the ratchet effect, I begin by assuming a linear functional form, which is standard in the educational literature.[19] This is done to develop intuition and make the identification strategy that follows more transparent — I later relax this assumption to explore whether allowing complementarities between inputs affects the results. 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 that there are average scores y_{scgt} and targets $\hat{y}_{scgt} \equiv \alpha y_{scg-1t-1}$ 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 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

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

where $F(\cdot)$ is the cdf of u, 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 is equivalent to $Pr[\sum_{g=1}^{G_c} u_{scgt} > \sum_{g=1}^{G_c} ((\alpha-\gamma) y_{scg-1t-1} - e_{scgt} - a_{scgt})]$, using equation (1).

3.2 Optimal Effort Levels

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

[19] For instance, Todd and Wolpin (2007) consider a series of linear specifications.

$$\frac{2d}{b}e_{scgt} = \begin{cases} f(\Pi_{sct}) + \delta(\gamma - \alpha)\sum_{i=0}^{G_c-g-1}\delta^i\gamma^i f(\Pi_{sc,t+1-i}) & \text{for } 1 \leq g < G_c \\ f(\Pi_{sct}) & \text{for } g = G_c \end{cases}$$

which cannot be used to solve for each effort level analytically. However, these first-order conditions can still be used to characterize the relationship between key parameters and the optimal effort levels.

Lemma 1 *Each optimal effort level is increasing in b and decreasing in the cost parameter d.*

The proof follows from the preceding conditions: Assuming all else is equal, intuitively speaking, a higher b causes the marginal benefit from effort to increase, while the marginal cost remains unchanged, leading the teacher to exert greater effort to bring the margins back into balance. To offer a concrete interpretation, if b is the bonus amount as a percentage of total teacher salary and the base non-performance-based salary of teachers increases with tenure (as is plausible), then the result can be interpreted as saying that the optimal effort level is decreasing in teacher experience. As for the quadratic cost parameter d, effort becomes less costly at the margin if it falls. Optimal effort must then adjust upward to restore equality between the marginal benefit and cost. This parameter d can be interpreted as a measure of how invested a teacher is in a particular teaching style. Less preparatory work should be required each year for a teacher who has taught the same curriculum or grade for a longer period of time. (In the language of the model, this corresponds to a more invested teacher possessing a higher d and exerting a lower level of effort.)

Imposing a steady state allows for the solution of effort in grade g to be easily expressed in terms of the effort in any other grade g'. In steady state, $e_{scgt} = e_{scg}$ and $\Pi_{sct} = \Pi_{sc}$, $\forall t$. Thus, the first-order conditions become

$$\frac{2d}{b}e_{scg} = \begin{cases} f(\Pi_{sc})\left[1 + \delta(\gamma - \alpha)\sum_{i=0}^{G_c-g-1}\delta^i\gamma^i\right] & \text{for } 1 \leq g < G_c \\ f(\Pi_{sc}) & \text{for } g = G_c \end{cases}$$

so that each e_{scgt} can be written in terms of a single base choice, such as e_{sc1}. The term contained in the square brackets for $1 \leq g < G_c$ is the distortion due to dynamic gaming, while $f(\Pi_{sc})$ represents the school-specific myopic incentives in the absence of a ratchet effect.[20]

[20]In the absence of dynamic target manipulation, different schools are expected to have different incentives to respond to the reform. In essence, teacher effort may matter more or less at the margin for receiving a bonus, leading to variation in the optimal response by teachers. This may be due to grade-to-grade differences in teacher ability and transitory shocks that revert to the mean in the following period.

Lemma 2 *Assuming that the high-powered target exceeds the growth rate of the score ($\alpha > \gamma$), steady-state effort is increasing in g.*

The proof is immediate from the preceding conditions. As the effort choice affects a larger number of future targets and the targets grow at a faster rate than the score ($\alpha > \gamma$), then teachers are increasingly penalized for exerting higher effort. Thus, it is optimal to select a lower level as the horizon increases (g is further away from the final grade offered G_c). For similar reasons, the converse is also true. That is, steady-state effort is 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 steady-state conditions. Therefore, I make an additional simplifying assumption that the incentive scheme is linear. In this 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. The conditions become

$$e_{cg} = \begin{cases} \frac{b}{2d}\left[1 + \delta(\gamma - \alpha)\sum_{i=0}^{G_c - g - 1} \delta^i \gamma^i\right] & \text{for } 1 \leq g < G_c \\ \frac{b}{2d} & \text{for } g = G_c \end{cases}.$$

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 ($y_{cg} > y_{c'g}, \forall\, g \in \mathcal{G}_c$).*

Proof For some $\kappa > 0$, 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. For the remainder of this proof, assume that $\delta > 0$, $\gamma > 0$ and $\alpha > \gamma$.

If $g = G$, then $e_{cG} = \frac{b}{2d}$ and $e_{c'G} = \frac{b}{2d}\left[1 + \delta(\gamma - \alpha)\sum_{i=0}^{\kappa - 1} \delta^i \gamma^i\right]$, which means that $e_{cG} > e_{c'G}$ from the stated assumptions.

If $1 \leq g < G$, then $e_{cg} = \frac{b}{2d}\left[1 + \delta(\gamma - \alpha)\sum_{i=0}^{G-g-1} \delta^i \gamma^i\right]$ and $e_{c'g} = \frac{b}{2d}\left[1 + \delta(\gamma - \alpha)\sum_{i=0}^{G+\kappa-g-1} \delta^i \gamma^i\right]$. 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 $e_{cg} = \frac{b}{2d}\left[\delta(\gamma - \alpha)\sum_{i=0}^{G-1} \delta^i \gamma^i\right]$ and $e_{c'g} = \frac{b}{2d}\left[\delta(\gamma - \alpha)\sum_{i=0}^{G+\kappa-1} \delta^i \gamma^i\right]$, 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, $e_{cg} > e_{c'g}, \forall\, g \in \mathcal{G}_c$.

Now, suppose that every student in a type c school begins grade $g = 1$ with an initial level of educational capital k_{c0}, and assume that this level is identical across school types, so that $k_{c0} = k_0, \forall\, c$. Also, assume that teacher ability by grade is identical across school types, so that $a_{cg} = a_g, \forall\, c$, and let the shock at the average school of each type c be zero ($u_{cg} = 0$). Thus, 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. ■

To interpret Proposition 1, consider the following example of a pair of average 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, respectively.[21] 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 three, four or five. The pattern of effort levels for K-5 and K-8 schools is illustrated in Figure 1. Thus, under the assumptions stated in Proposition 1, the test score for any particular shared grade is predicted to be higher at the K-5 school when compared to the K-8 school. An example of such a pattern in scores is shown in Figure 2; 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, $\forall\ g \in \mathcal{G}_c$.*

Proof 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 than 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, $\forall\, g \in \mathcal{G}_c$. ■

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

[21] Recall that only grades with high-powered incentives attached are relevant to the discussion and that grades three and up satisfy this criterion in North Carolina.

should be greatest for grade five — this result is reflected in Figure 2. The grade five result is the main prediction to be tested empirically.

Proposition 3 *As an analog to Lemma 1, the positive score differential between two schools of different types is increasing in b and decreasing in the cost parameter d.*

Proof Under the same assumptions used in the proof for Proposition 1, the disparity in score is equal to the disparity in cumulative effort. That is, $y_{cg} - y_{c'g} = \sum_{i=1}^{g} \gamma^{g-i}(e_{ci} - e_{c'i})$. From the proofs of Propositions 1 and 2, $e_{cg} - e_{c'g} = \frac{b}{2d}\left[\delta(\alpha-\gamma)\rho_{\kappa g}\right] > 0$. Since $e_{cg} - e_{c'g}$ is increasing in $\frac{b}{2d}$, $\forall\ g \in \mathcal{G}_c$, it follows that $y_{cg} - y_{c'g}$ is increasing in $\frac{b}{2d}$, $\forall\ g \in \mathcal{G}_c$. ■

Under the same interpretation as given for Lemma 1, there will be a greater distortion between two configurations for teachers with less experience (larger b) and those who have invested less in teaching a specific curriculum (smaller d).

Proposition 4 *Even without assuming a steady state or a linear incentive scheme, dynamic distortions are eliminated if the planner sets the target coefficient α to be the same as the growth rate γ.*

The proof is immediate from any of the first-order conditions. Under the most general conditions, if $\alpha = \gamma$, then optimal effort is equal to $f(\Pi_{sct})$, $\forall g \in \mathcal{G}_c$, meaning that effort is identical across grades and the ratchet effect disappears. By matching the growth rate with the target coefficient, the scheme no longer punishes teachers in the future for exerting higher effort today. Instead, an increase in the next-period target from greater contemporaneous effort is exactly met by an equal increase in the next-period score. Figure 3 shows what the resulting scores by grade and grade span might look under such a scenario.

3.3 Extensions Under a Linear Scheme and Linear Technology

The preceding linear model is readily generalized in two dimensions. First, incorporating grade-specific growth rates allows for the possibility that students in earlier grades experience greater or lesser growth independent of any new inputs. Second, transitory processes can be introduced — a dimension that is particularly interesting and relevant to the literature on 'teaching to the test.'

Thus far, I have treated effort as an input that persists as fully as a student's underlying educational capital. However, if a teacher focuses on teaching to a specific test in a given year, this component of her effort may not readily transfer to the following year through

her students' educational capital. I now formalize these ideas and compare them briefly to the simpler model already developed. When growth rates are grade-specific, the production technology becomes

$$y_{scgt} = \gamma_g y_{scg-1t-1} + e_{scgt} + a_{scgt} + u_{scgt}\,, \tag{3}$$

and the payoff function for school s is given by equation (2), with the slight exceptions of using the grade-specific quantities γ_g and α_g in place of γ and α, respectively. Given these changes, the first-order conditions under a linear incentive scheme are

$$e_{cg} = \begin{cases} \frac{b}{2d} & \text{for } g = G_c \\ \frac{b}{2d}\big[1 + \delta(\gamma_{g+1} - \alpha_{g+1})\big] + \delta\gamma_{g+1}\big(e_{cg+1} - \frac{b}{2d}\big) & \text{for } 1 \leq g < G_c \end{cases},$$

where the conditions are defined recursively for $1 \leq g < G_c$.

Proposition 5 *If $\alpha_g > \gamma_g$, $\forall\, g \in \mathcal{G}_c$, then teacher effort is increasing in the grade g.*

Proof For $g = G_c - 1$, the first-order condition is $e_{G_c-1} = \frac{b}{2d}\big[1 + \delta(\gamma_{G_c} - \alpha_{G_c})\big]$, since $e_{G_c} = \frac{b}{2d}$. Therefore, $\alpha_{G_c} > \gamma_{G_c}$ implies that $e_{G_c} > e_{G_c-1}$.

For $g = G_c - 2$, the condition is $e_{G_c-2} = \frac{b}{2d}\big[1 + \delta(\gamma_{G_c-1} - \alpha_{G_c-1})\big] + \delta\gamma_{G_c-1}\big(e_{G_c-1} - \frac{b}{2d}\big)$ or $e_{G_c-2} = \frac{b}{2d}\big[1 + \delta(\gamma_{G_c-1} - \alpha_{G_c-1}) + \delta^2\gamma_{G_c-1}(\gamma_{G_c} - \alpha_{G_c})\big]$. Using $\delta < 1$, $\gamma_g < 1$ and $\alpha_g > \gamma_g$, $\forall\, g \in \mathcal{G}_c$, it must be the case that $\delta^2\gamma_{G_c-1}(\gamma_{G_c} - \alpha_{G_c}) < \delta(\gamma_{G_c} - \alpha_{G_c})$ and $\delta(\gamma_{G_c-1} - \alpha_{G_c-1}) < 0$. Therefore, $e_{G_c-1} > e_{G_c-2}$.

Similar reasoning applies for $1 \leq g < G_c - 2$. ■

Thus, the key intuition developed under the grade-invariant growth model continues to hold, meaning that the result is not an artifact of the parameter restriction.

With respect to modeling transitory processes, I allow for the teacher inputs and the shock to persist into the future at a rate of $\omega\gamma_g$, where $0 < \omega < 1$, rather than the rate γ_g for the existing stock of educational capital. The production technology now becomes

$$y_{cgt} = \gamma_g y_{cg-1t-1} + \gamma_g(\omega - 1)(e_{cg-1t-1} + a_{cg-1t-1} + u_{cg-1t-1}) + e_{cgt} + a_{cgt} + u_{cgt}\,, \tag{4}$$

where $y_{cg-1t-1} - e_{cg-1t-1} - a_{cg-1t-1} - u_{cg-1t-1}$ and $e_{cg-1t-1} + a_{cg-1t-1} + u_{cg-1t-1}$ evolve at rate γ_g and $\omega\gamma_g < \gamma_g$, respectively.

The corresponding first-order conditions with respect to effort yield $e_{G_c} = \frac{b}{2d}$ for the effort level in the final grade served, $e_{G_c-1} = \frac{b}{2d}\big[1 + \delta(\omega\gamma_{G_c} - \alpha_{G_c})\big]$ for the effort level in the second from last grade served, and

$$e_g = \frac{b}{2d}\Big[1 + \delta(\omega\gamma_{g+1} - \alpha_{g+1}) + \delta\omega \sum_{i=1}^{G_c-g-1} \delta^i(\gamma_{g+1+i} - \alpha_{g+1+i}) \prod_{j=1}^{i} \gamma_{g+j}\Big]$$

for all other grades $1 \leq g \leq G_c - 2$.

Proposition 6 *If $\alpha_g > \gamma_g$, $\forall\, g \in \mathcal{G}_c$, then the dynamic distortion for grades $G_c - 1$ and $G_c - 2$ increases as teacher inputs become less persistent (ω decreases). The result holds for $g < G_c - 2$ as long as the difference between α_g and γ_g is sufficiently small, $\forall\, g \in \mathcal{G}_c$.*

Proof From the first-order conditions, the distortion in effort for grade $g = G_c - 1$ is $\omega\gamma_{G_c-1} - \alpha_{G_c-1} < 0$, which becomes less negative as ω rises $\big(\frac{\partial(\omega\gamma_g - \alpha_g)}{\partial\omega} = \gamma_g > 0\big)$, meaning that the disparity between effort in grade G_c and $G_c - 1$ is magnified as ω falls.

For $g = G_c - 2$, the distortion is $\frac{b}{2d}\big[\delta(\omega\gamma_{G_c-1} - \alpha_{G_c-1}) + \delta^2\omega\gamma_{G_c-1}(\gamma_{G_c} - \alpha_{G_c})\big]$, when compared to effort in grade G_c. Its derivative with respect to ω is then $\frac{b}{2d}\delta\gamma_{G_c-1}\big[1 + \delta(\gamma_{G_c} - \alpha_{G_c})\big]$, which is positive since $\alpha_g < 1$, $\gamma_g < 1$ and $\alpha_g > \gamma_g$, $\forall\, g \in \mathcal{G}_c$.

For $1 \leq g < G_c - 2$, the distortion is $\frac{b}{2d}\big[\delta(\omega\gamma_{g+1} - \alpha_{g+1}) + \delta\omega \sum_{i=1}^{G_c-g-1} \delta^i(\gamma_{g+1+i} - \alpha_{g+1+i}) \prod_{j=1}^{i} \gamma_{g+j}\big]$. Its derivative with respect to ω is $\frac{b}{2d}\delta\gamma_{g+1}\big[1 + \sum_{i=1}^{G_c-g-1} \delta^i(\gamma_{g+1+i} - \alpha_{g+1+i}) \prod_{j=2}^{i} \gamma_{g+j}\big]$, which is positive if $\sum_{i=1}^{G_c-g-1} \delta^i(\alpha_{g+1+i} - \gamma_{g+1+i}) \prod_{j=2}^{i} \gamma_{g+j} < 1$.

This condition holds if $\alpha_g - \gamma_g$ is sufficiently small, $\forall\, g \in \mathcal{G}_c$. ■

Given that a falling ω is equivalent to greater 'teaching to the test,' this proposition means that such 'static' gaming of the system actually magnifies the dynamic distortion in effort. The next proposition is an analog to Proposition 4.

Proposition 7 *Dynamic distortions in the presence of transitory effort can be eliminated if the planner has the flexibility to choose grade-specific target coefficients α_g.*

To eliminate distortions, the final-grade target coefficient should be $\alpha^*_{G_c} = \omega\gamma_{G_c}$, which is readily apparent from the expression for effort in grade $G_c - 1$. The second-from-last grade target coefficient should be $\alpha^*_{G_c-1} = \omega\gamma_{G_c-1} + \delta\omega\gamma_{G_c-1}\gamma_{G_c}(1 - \omega)$ and is calculated by substituting $\alpha^*_{G_c}$ into the expression for e_{G_c-2}, equating it to the expression for e_{G_c}, and solving for α_{G_c-1}. Coefficients for lower grades served are calculated in the same way, but are omitted here for brevity.

3.4 Complementarity in Production

The simple linear production technology defined in equation (1) is in line with the existing education literature, but potentially ignores important features of the learning process. Chief among them is the possibility that teachers exert effort differentially by student ability. For instance, teacher effort may have a greater effect on the scores of students who are of higher ability. Conversely, the greatest gains in learning per unit of effort may be realized from students who struggle most. In either case, nonlinear interactions in the production process may nontrivially affect how dynamic distortions manifest themselves. Consider the production technology

$$y_{scgt} = \gamma y_{scg-1t-1} + \theta e_{scgt} y_{scg-1t-1} + e_{scgt} + a_{scgt} + u_{scgt}\,, \tag{5}$$

where θ is the interaction parameter determining whether teacher effort and student ability, as proxied by the prior score $y_{scg-1t-1}$, are complements or substitutes in production. Conveniently, $\theta = 0$ recovers the simplified linear process presented in equation (1), making it a special case of this more general formulation.

Given that all other aspects of the theoretical environment remain unchanged and assuming a linear incentive scheme to allow for analytical solutions, the optimal effort levels are governed by a set of first-order conditions that grow increasingly complex as the grade g becomes more distant from the last grade served G_c. Defining $B \equiv b/(2d)$, the simplest condition is for the last grade served and is given by

$$e_{scG_ct} = B(1 + \theta y_{scG_c-1t-1})\,, \tag{6}$$

which features no distortion, just as in the linear production case, but does scale according to the prior score and parameter θ. The condition for the second-from-last grade served $G_c - 1$ is given, for illustration, by

$$e_{scG_c-1t} = \frac{B(1 + \theta y_{scG_c-2t-1})(1 + \delta[\gamma - \alpha + 2B\theta(1 + \gamma\theta y_{scG_c-2t-1})])}{1 - 2\delta B^2\theta^2(1 + \theta y_{scG_c-2t-1})^2}\,, \tag{7}$$

where the distortion is effectively $\delta[\gamma - \alpha + 2B\theta(1 + \gamma\theta y_{scG_c-2t-1})]$.[22] Thus, an interaction between teacher effort and student ability not only causes optimal effort to scale with the prior score, but also affects the magnitude of the dynamic distortion. The expressions for effort in grades $G_c - 2$ and lower are more involved and are omitted here, but we can proceed analytically.

[22]The denominator of equation (7) is of second-order importance when comparing distortions.

Proposition 8 *For $\alpha > \gamma$, $\theta \ll \gamma$ and identical prior scores, teacher effort is greater in grade G_c than in grade $G_c - 1$. For a given growth rate γ, this distortion is magnified compared to the linear production technology result if $\theta < 0$ and is attenuated if $\theta > 0$.*

The proof is immediate from conditions (6) and (7). It is important to note that the dynamic distortion can no longer be eliminated by equating α and γ as in Proposition 4. Instead, the best that can be done under a linear target is to eliminate the average distortion by setting $\alpha = \gamma + 2B\theta(1 + \gamma\theta\bar{y}_{cG_c-2t-1})$, where $\bar{y}_{cG_c-2t-1}$ is the average prior score for school configuration c. This state of affairs is entirely due to the fact that the distortion now contains a nonlinear component. Therefore, the only way to fully compensate for it is to employ a more complicated nonlinear target, which is beyond the scope of the current treatment.

3.5 Extensions

There are several ways in which the model can be extended. First, the linear incentive scheme assumption can be relaxed to explore the implications of allowing for the type of nonlinear threshold-based scheme used in practice. This is one focus of my ongoing research. In addition, the model does not yet differentiate between rival mechanisms for principals to respond to the scheme, initially focusing exclusively on the monitoring and coordination of teacher effort. In related work, I allow for the additional possibility that principals re-allocate teachers across classrooms to maximize their school's payoff. The existence of this alternative channel should not affect the empirical identification strategy discussed below, since such behavior is predicted to have observationally equivalent effects on student scores as in the current formulation focusing on teacher effort.

4 The North Carolina Accountability Reform

In this section, I first describe the broad evolution of accountability in North Carolina. Then, given that my identification strategy is based on the accountability scheme that North Carolina adopted in 1996, I describe it in greater detail. In particular, I establish the following features of the scheme, captured in a stylized way by the preceding theory: school-level targets depend on the prior scores of students, the average of grade-specific targets must be satisfied in order to receive a bonus, and the grade span determines the number of such targets included in the average.

4.1 A Brief History

As Fabrizio (2006) notes, standardized testing across the state and reporting of the resulting scores to the public at the district level were introduced through the passage of the School Accountability Act in 1989. By 1991, district report cards were published which corrected for socioeconomic measures, such as parental education.[23] This low-stakes accountability environment lasted until 1993. It was in that year that more coherent norm-referenced end-of-grade reading and mathematics tests were established for students in grades three through eight. These formed the backbone of the high-stakes accountability system that followed.

Called the ABCs of Public Education (which stands for strong Accountability, teaching the Basics and focusing on local Control), the legislation for the initial incarnation of the current accountability reform in North Carolina was ratified in June of 1996 and implemented beginning in the 1996-97 school year for all schools in the state serving students in kindergarten through grade eight.[24] This occurred after a pilot phase in the 1995-96 school year 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 ABCs reform.

There was strong bipartisan political support for the accountability reform. According to Fabrizio (2006), who synthesizes the data collected from extensive interviews of people knowledgeable about the ABCs history, the reform was spearheaded by democratic Governor James Hunt, who held office from 1993 to 2001 and who was a proponent of standards-based education reform, and also Dr. Jay Robinson, a former chairman of the State Board of Education. Other stakeholders instrumental in the formation and passage of the reform included leaders in the private sector and the North Carolina Association of Educators.[25]

In the decade following passage of the ABCs reform, some minor modifications to the program were implemented, none of which significantly affect the dynamic gaming analysis in this paper.[26] A more substantial overhaul of the way incentive targets are calculated was

[23] See Ladd (2001) for details.

[24] Basic information about the North Carolina ABCs is found in an electronic brochure (source: *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).

[25] The support of educators is interesting in light of the fact that there is no collective bargaining in the state. Passed in 1959, General Statute 95-98 voids any contract between a labor union and the State of North Carolina or municipal governments — see:
http://www.ncga.state.nc.us/gascripts/statutes/StatutesTOC.pl?Chapter=0095

[26] Although not the focus of the current study, the program was expanded to include high schools in the 1997-98 school year. Second editions of the standardized tests for elementary students in math and reading

carried out in the 2005-06 school year, which explains why my investigation into dynamic gaming does not extend beyond 2005. However, the salient features of the reform 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. At the time of the 2005-06 change, Fabrizio (2006) reports that the program had dispensed over $870 million in pecuniary payments to educators throughout its existence, according to a source at the Department of Public Instruction in North Carolina.

Alongside these adjustments to the ABCs reform, a low-stakes student accountability reform was introduced in the 2000-01 school year, requiring students in grade five to score above a specific proficiency level in reading and mathematics in order to be promoted to sixth grade.[27] The program was expanded in the 2001-02 school year to include students in grades three and eight.[28] Provisions of the federal No Child Left Behind Act were then implemented in the 2002-03 school year. In each of the aforementioned cases, the associated targets do not account for student, teacher or school heterogeneity, likely leading the incentives to be pronounced for only a small subset of students or teachers. To the extent that grade-specific student accountability may affect my empirical results, I address these concerns when discussing the robustness of my econometric strategy.

4.2 The Reform in Detail

Under the 1996 reform, students in grades three through eight are required to write standardized tests in reading and mathematics at the end of each school year. Using this information, subject-specific growth targets are calculated for each student using his or her prior performance in each subject. The targets are then aggregated to the school level to form expected growth scores for each school.[29] The grade-specific expected growth targets for reading and math are given by the following formulae:

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

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

were introduced in the 2000-01 and 2002-03 school years, respectively, with equating studies being done to ensure comparability between first and second editions. Also in the 2000-01 school year, the reduced-form target coefficients of the scheme were updated for grade three.

[27] See Cooley (2010) for an analysis of the initial grade five reform.

[28] Details on the student-focused reform for all affected grades can be found online — see: *http://www.ncpublicschools.org/docs/promotionstandards/policy/overviewstandards.pdf*

[29] 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.

where $\Delta\widehat{r}_{gst} \equiv \widehat{r}_{gst} - r_{sg-1t-1}$, $\Delta\widehat{m}_{gst} \equiv \widehat{m}_{gst} - m_{sg-1t-1}$, r_{gst} and m_{gst} are the average reading and math scores for school s in 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, 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. These expected growth targets (or gains) were calculated for every grade in a school for each year beginning with the 1996-97 school year.[30]

The first component of each expected gain ($\hat{\alpha}_0$ or $\hat{\beta}_0$) is the mean expected gain across all schools in the state. 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. To be clear, consider 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 toward the state average over time.[31]

In each year, the expected gains are used to form a composite score for each school by taking the difference between the school's realized growth Δy_{st} and expected growth $\Delta\widehat{y}_{st}$ in each subject ($y \in \{r, m\}$) and dividing this by the standard deviation of all scores for that subject in the state (σ_t^y). The resulting standardized composites for each subject are then combined to form the main composite.[32] This composite is used to determine whether educators at a school receive a bonus. If the main composite for their school is positive, then the principal and all teachers 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.[33]

As mentioned, the expected growth target coefficients are given. They are estimated from score data in the 1992-93 and 1993-94 school years by regressing the actual score gain in the 1993-94 school year on the ability and mean reversion proxies for each subject, namely the

[30] Although the expected gains for each grade at a school are combined to determine whether educators will receive a bonus, I suppress grade subscripts for the remainder of this section to simplify exposition.

[31] 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 are out of their control. On this basis, it is desirable to correct for mean-reverting processes. While the North Carolina approach can only adjust for transitory phenomena that affect subjects differentially, it is noteworthy that policymakers made an effort to address the problem of period-by-period noise.

[32] The main composite is actually composed of reading and math composites for each grade at a school. For the purposes of this institutional discussion, I continue to abstract from this fact, but it will become very important for the empirical analysis that follows.

[33] A teacher with 13 years of experience and a Bachelor's degree made about \$30,000 in the 1997-98 school year. Thus, \$1,500 is approximately equal to 5% of yearly pay or 60% of monthly pay.

combined prior score for ability and the subject-specific prior score for reversion. Specifically, defining $\tilde{r}_{st} \equiv r_{st} - \bar{r}_t$ and $\tilde{m}_{st} \equiv m_{st} - \bar{m}_t$ as the demeaned reading and math scores for school s in year t,[34] the actual gain in reading $\Delta r_{s,94}$ is regressed on $\tilde{r}_{s,93} + \tilde{m}_{s,93}$ and $\tilde{r}_{s,93}$ to obtain $\hat{\alpha}_0$, $\hat{\alpha}_1$ and $\hat{\alpha}_2$, and the actual gain in mathematics $\Delta m_{s,94}$ is regressed on $\tilde{r}_{s,93} + \tilde{m}_{s,93}$ and $\tilde{m}_{s,93}$ to obtain $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$.[35] Once estimated, these coefficients are used in all subsequent years when calculating expected gain targets. As such, they are treated as fixed in my analysis. It is also important to note that the state means ($\bar{r}_t$ and $\bar{m}_t$) and standard deviations (σ_t^r and σ_t^m) are not calculated contemporaneously with the expected gain, but rather are calculated using score data in the 1994-95 school year and fixed at that value for future years.

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 a number of 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. For instance, the differential ability of the prior teacher and/or school may contribute. If the prior teacher is much more able than the current one, the target will be overly difficult for the latter teacher, meaning that it will only be exceeded with extraordinarily high effort. To the extent that teachers have an effect on both scores of their students, this further complicates inference of student ability from the combined prior score. Allowing for transitory effects, as the scheme does by attempting to correct for mean-reverting tendencies, additional distortions in the target become likely. Any temporary effects that influence both reading and math in a given year will be mistakenly attributed as permanent effects under the North Carolina scheme. Thus, the attainability of the target may potentially depend on random shocks as well as on partially transitory teacher effort, the latter of which is expected to vary over time if teachers are held accountable in the prior period. These are undesirable aspects of the reform, since teachers are then held accountable for an outcome that they do not fully control.

While the 1996 North Carolina accountability reform is not without its flaws, there are several elements that make it suitable for detecting evidence of dynamic gaming behavior. It

[34] The year t refers to the school year ending in that year. For instance, $t = 94$ refers to the 1993-94 school year.

[35] 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.

features a high-powered school-level reward scheme that conditions targets on 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. Indeed, survey evidence lends credence to this idea.[36] Moreover, the program is long-standing, so that any dynamic distortions would have had a chance to manifest themselves. As such, the incentive environment is conducive to testing my theoretical predictions. I now turn to describing the data underlying my empirical analysis.

5 Data and Descriptive Statistics

5.1 Description

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 detailed information on North Carolina students, teachers and schools for the years 1994 through 2005.[37] Given that the accountability reform took effect in 1997, I refer to 1994, 1995 and 1996 as pre-reform years, and 1997 through 2005 as post-reform years.[38] The data set contains yearly standardized test scores for each student in mathematics and reading from grades two to eight.[39] These scores are comparable across time and grades through the use of a developmental scale.[40] Using this scale and unique

[36]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.

[37]For a graphical representation of the available data, see Appendix A.1. The student-level data extends from 1993 to 2008. However, the reform was substantially altered in 2006 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.

[38]The reform was implemented as a pilot program in 1996 for ten school districts, consisting of 63 schools or approximately 4 percent of schools in North Carolina. These schools are more rural and are slightly more likely to be K-6 or K-8 than the state average. As an alternative, defining 1994 and 1995 as the pre-reform period does not affect my results.

[39]'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.

[40]The developmental scale is formed from the number of correctly answered questions on the standardized test. By design, each point of the developmental scale is meant to measure the same amount of learning, so that a child whose score increases from 300 to 301 corresponds to the same amount of learning as another child realizing an increase from 310 to 311. Moreover, the same comparison holds true across grades, meaning that a child who realizes identical growth on the developmental scale in two consecutive grades is interpreted as learning equal amounts in each year.

encrypted identifiers, the progress of individual students can be tracked over their educational careers. 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 a principal has been in charge of a school, the number of classes by grade offered by a school, and — especially relevant for this study — each school's grade configuration.

Descriptive statistics for the variables of interest are presented 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. As expected from the developmental scale, the mean combined math and reading score is increasing in the grade. In addition, with the exception of the gain from grade six to seven, the rate of growth is decreasing in the grade, so that students gain the most in grade four, followed closely by grade five. With respect to non-score data, students with parents who possess a high school diploma and no post-secondary education account for 42% of the sample, while those who have not obtained a high school degree make up 10%. Parents with a diploma from a trade school or community college account for a further 20% of the sample, and 28% of parents have been granted a 4-year college or graduate degree. Nearly two-thirds of North Carolina students are white in the sample, while slightly less than 30% are black, which is significantly higher than the national average and also higher than the state average for North Carolina.[41] In the data set, the average teacher has about 13 years of teaching experience, the school-level average percentage of students qualifying for a free lunch is 38%, and the average number of classes in grades three through five at all school configurations is 3.5.

As for 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 and 489, 71

[41] According to a 2009 estimate by the U.S. Census Bureau, approximately 13% and 22% of the U.S. and North Carolina population, respectively, are identified as being black (source: *http://quickfacts.census.gov/qfd/states/37000.html*).

and 24, respectively, with the additional restriction that the school is observed in all pre- and post-reform years of the sample. The strong decline in K-6 schools between the least and most restrictive samples can be attributed to the fact that many of those open in the pre-reform period close or switch to a K-5 configuration early in the post-reform period. Nevertheless, 264 school-year observations for K-6 schools remain even under the most restrictive subsample, which is sufficient for the purpose of analyzing gaming behavior across grade spans.

Upon cursory inspection of the data, K-8 and K-6 schools are observably different from their K-5 counterparts. For instance, K-8 and K-6 schools are disproportionately located in rural areas, while K-5 schools tend to be found more in urban and suburban areas.[42] This is important since students at rural schools tend to be more economically disadvantaged, as measured by greater participation in the free lunch program, tend to have parents with lower educational attainment, and are less likely to be black. Given the over-representation of the two comparison configurations (K-8 and K-6) in rural areas, controlling for the school's locale in the analysis is therefore likely to be important.

5.2 The Impact of the Reform

As a starting point, it is instructive to see which patterns emerge in the raw data. There are two features that are particularly interesting. The first relates to whether the reform had a positive effect on scores overall. That is, did it do what it was supposed to do? The second concerns whether the reform had stronger effects in certain school configurations. Both of these questions can be addressed in a clear way using distributional plots.

Figures 4a and 4b show density plots of first-differenced student scores by grade, for grades two through five, using raw scores and scores that are adjusted for observable characteristics, respectively.[43] The first thing to note is that the mean of each distribution is positive, reflecting the fact that the average post-reform score is greater than its pre-reform counterpart. This evidence is in line with the notion that the accountability reform improved overall scores. Another interesting feature is that the growth in scores is monotonically increasing in the grade, which is precisely 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,

[42] For the full sample, approximately 396 K-5 schools, 87 K-8 schools and 71 K-6 schools are located in rural areas. For the subsample of schools that do not switch and those observed in all pre- and post-reform years, the counts are 297, 69 and 30, and 226, 62 and 20, respectively.

[43] 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.

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.

Decomposing the grade five score by school configuration is also suggestive. Figure 5 plots the density and means (given by the vertical lines) of the first-differenced grade 5 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. Due to fewer observations, the underlying distributions for K-6 and K-8 schools are not as smooth as the equivalent for K-5 schools. Although it seems as if K-6 schools have a lower mean than K-8 schools, the opposite cannot be statistically ruled out as the associated confidence intervals are both much wider than for K-5 schools. In any case, the main comparisons of interest are between K-5 and K-6, or K-5 and K-8 schools. With this suggestive evidence in hand, I now set out my basic econometric strategy to test formally for ratchet effects.

6 Reduced-Form Analysis

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 are otherwise identical. Such a strong condition — that grade spans are exogenous — is unlikely to be satisfied in practice. Therefore, I develop a reduced-form strategy to control for unobserved differences across schools and present the associated results. I subsequently carry out a falsification exercise and explore the results along various dimensions.

6.1 Econometric Strategy

Owing to a variety of historical factors, the popularity and adoption of different elementary school grade configurations has waxed and waned over time, potentially leading such config-

urations to be non-randomly represented in the current population of schools.[44] 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 differences. 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

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

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. On an observable basis, 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

[44] 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, transitionary 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 (2001), Hanushek *et al.* (2004) and Rockoff and Lockwood (2010), also suggested that a higher number of school transitions was deleterious to student development.

$$y_{scgt} = X'_{sgt}\beta + \sum_{c=1}^{C} \sum_{g \in \mathcal{G}_c} (\phi_{c,g,pre} + \phi_{c,g,post}) + \varepsilon_{scgt} \tag{8}$$

where each ϕ is an interacted indicator variable that adjusts the score for every combination of grade, school type and period.[45] Effectively, 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 (8), 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

$$\Phi_{\text{K5}-\text{K8},g,post-pre} = (\phi_{\text{K5},g,post} - \phi_{\text{K5},g,pre}) - (\phi_{\text{K8},g,post} - \phi_{\text{K8},g,pre}) \,. \tag{9}$$

In the event that unobserved trends are common across grade configurations, a finding of $\Phi_{\text{K5}-\text{K8},g,post-pre} > 0$ is interpreted as satisfying the criterion for dynamic gaming behavior as in Proposition 1.

Despite the merits of the proposed difference-in-differences strategy that includes controls, differentially trending unobservables may engender a nontrivial source of bias in the estimates, the direction of which is not obvious. Demand-side sorting by households or teachers across schools of different grade configuration, due to unobserved evolving differences in neighborhood or workplace amenities,[46] and supply-side changes in the distribution of school configurations over time are potential areas for concern. With respect to the latter supply-side issues, North Carolina policymakers increasingly shifted toward the K-5/6-8 model during the post-reform period.[47] If the schools were systematically selected for such a transition on the basis of unobserved determinants of performance, bias would result.[48]

[45] Neither allowing control coefficients to vary by grade (β_g) (which is a prerequisite for structurally estimating the model with transitory effort) nor including school-level fixed effects appreciably alters the difference-in-differences results.

[46] K-6 and K-8 schools are predominantly found in rural areas. If economic opportunities in these areas disproportionately decrease for low-ability households, for example, such families may increasingly migrate to urban centers, biasing difference-in-differences estimates downward. More problematic upward bias would result if shifting economic conditions had a larger effect on high-ability households or were geographically reversed. In addition to changing salary differentials across school districts, such evolving conditions might also lead to teacher sorting and associated bias.

[47] Between 1995 and 2005, there was a 27% and 79% decline in the number of K-8 and K-6 schools, respectively, with an offsetting 56% increase in the number of K-5 schools.

[48] For instance, underperformers might be chosen first due to having less institutional resistance. If such schools tend to be located in disadvantaged neighborhoods, then average student ability would rise for K-8 and K-6 schools and fall for K-5 schools after the transition, leading to downward-biased estimates. It could also be the case that high-performing schools would prefer to undertake the transition if the associated benefits

One approach for addressing the supply-side issues that have been raised 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 subsection. However, even after restricting the analysis, selection bias may remain due to the competitive effects of switching schools on non-switching ones, assuming schools compete with each other locally.[49] Analyzing student migration between switching and non-switching schools might help shed light on the direction of such bias, but as with demand-side effects which are unlikely to be fully captured by observed characteristics, there is a more robust way to deal with this issue.

Recall that the basic reduced-form strategy entails taking the difference-in-differences of scores between the pre- and post-reform period and between two configurations. Given that this is done for every grade that is shared by the configurations, a triple difference can be formed using the difference between such estimates for any two grades. For instance, the estimate comparing grade four and five scores between K-5 and K-8 schools is

$$\Phi_{\text{K5}-\text{K8},5-4,post-pre} = \Phi_{\text{K5}-\text{K8},5,post-pre} - \Phi_{\text{K5}-\text{K8},4,post-pre}\,. \tag{10}$$

Such an analysis not only 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 does not affect scores differentially by configuration and grade, then remaining demand- or supply-side selection bias is addressed by the triple-differences approach. For this reason, it is the preferred econometric strategy. A finding of $\Phi_{\text{K5}-\text{K8},5-4,post-pre} > 0$ 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.

offset the costs, which would bias the estimates upward. However, the average pre-reform grade five score for K-8 schools that transitioned to a new configuration and those that did not is 309.0 and 310.2, respectively (a difference of 7.1% of a standard deviation in the grade five score), providing evidence in line with the downward bias story.

[49] To see why, consider the example of a district with two K-8 schools, one of which is underperforming, and the other, high performing. Let the underperforming one convert to a K-5 school. If such a configuration is more desirable than a K-8 one, then the new school may attract some higher ability students from the K-8 school that remains in the non-switching sample, resulting in lower average student ability at the school. If this was the case, then the estimates would be biased upward. Conversely, downward bias would result if the newly converted K-5 school were perceived as being less desirable.

6.2 Results

Figures 4a, 4b and 5 already provided preliminary evidence consistent with dynamic gaming. I now analyze these effects in a more econometrically rigorous way. In particular, I estimate equation (8) under three different specifications, dictated by the components of the control vector X_{sgt}. These specifications are given in Table 2, 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.[50]

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 labelled as being exceptional. For specification (3) in particular, relative to a student with a parent who has not finished high school, a child with a parent whose highest educational attainment is a high school diploma has a predicted score that is approximately 6.8 developmental scale points higher, while a student with a parent who possesses a four-year college degree extends this gain by a further 8.0 points. With respect to ethnicity, a black student is predicted to have a score that is 8.4 points lower than a non-black student. These are large differences, as the standard deviation of the grade five score reported in Table 1 is 16.9 developmental points. The score is also linked positively to students attending a school with a lower free lunch participation rate and those with teachers who scored higher on their licensing test. In the case of free lunch, a decrease of one standard deviation in the participation rate at a school is associated with an increase of one developmental point for students at that school.

For specifications (1) through (3) in Table 2 and grades three through five, I transform the relevant fixed effects from equation (8) into first-difference, difference-in-differences and triple-differences estimates, as in equations (9) and (10). The results for K-5 and K-8 schools, and K-5 and K-6 schools are reported in Table 3. In every case, the difference between pre- and post-reform scores for a specific configuration is positive and significant, which is 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

[50]Specifications (2) and (3) are separately presented, since teacher-level variables are not observed for the first pre-reform year (1994). While it seems reasonable to control for teacher characteristics, one might be worried that omitting one of the three pre-reform years would substantially alter the results. I report both specifications to show that this is not the case.

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. This highlights the fact that score growth increases with the grade regardless of the school's grade configuration.[51]

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 3 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.[52] 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 insignificantly positive for the comparison between grades four and three when including controls. These results are in line with 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 is between 8.6% and 9.1% of a standard deviation in the grade five score. For the more rigorous triple differences, 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% of a standard deviation in the grade five score).

It is informative to place these results in context. Using the figures above, the score differential between a student with a parent who graduated high school compared to one who did not is 40.2% of a standard deviation in the grade five score, while 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% of a standard deviation in the grade five score. Thus, the dynamic distortion between K-5 and K-8 schools is considerably weaker than the parental education effect and only slightly weaker than the effect of reducing the proportion of students eligible for free lunches.

In contrast to the comparison between K-5 and K-8 schools, Table 3 shows that the difference-in-differences estimates for K-5 and K-6 schools using specification (1) are positive

[51]From Table 1, the standard deviation of the score in grade four and three is 18.3 and 18.8 points, respectively, which is actually higher than the value for grade five (16.9 points). Thus, adjusting for variation in scores, the grade three and four gains are even smaller relative to those in grade five.

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

and statistically significant for all grades. This pattern continues to hold when including controls for differentially trending observables. Across all specifications, the grade five distortion accounts for between 1.71 and 2.84 developmental scale points. Although these magnitudes are larger than for the comparison between K-5 and K-8 schools, this relationship inverts when considering the triple-differences estimates. These preferred estimates are significant for the comparison between grades five and four when including controls and range between 0.66 and 0.95 scale points (or between 3.9% and 5.6% of a standard deviation in the grade five score).[53]

In Table 4, I compare difference-in-differences and triple-differences results for the full sample of schools to those 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.[54] Thus, the sign and significance of the difference-in-differences and triple-differences estimates are consistent with the dynamic gaming hypothesis, under both the full sample and the restricted subsample.

6.3 Falsification Exercise

The primary remaining threats to validity concern the implementation of other educational reforms during the period of analysis. For instance, North Carolina began allowing charter schools to compete with conventional public schools in 1998.[55] If charter schools cause the average scores of neighboring public schools to rise through increased competition, and charter schools are introduced into districts non-randomly according to school configuration, then this reform could cause bias in the estimated dynamic distortion.[56]

[53]Despite the standard deviation of scores across K-8 schools being slightly lower than the analogous measure for K-6 schools, K-6 schools are observed nearly twice as often in the pre-reform period as K-8 schools, while K-8 schools are slightly more numerous in the post-reform period. Therefore, given the critical role that the number of pre-reform observations plays, the comparisons between K-5 and K-6 schools tend to be more precise than those involving K-5 and K-8 schools.

[54]The analogous estimates for the K-5 and K-6 comparison are now insignificant and decrease somewhat to between 0.83 and 0.92 scale points. The insignificance is due to the increase in standard errors which occurs as a result of many more K-6 schools converting into K-5 schools than K-8 into K-5 during the period of study.

[55]From Bifulco and Ladd (2004), 27 charter schools began operating in 1998, with the number growing to 67 by 2002.

[56]However, note that for a subset of the North Carolina data used in this study, Bifulco and Ladd (2004) find that the effect of charter schools on public schools is negligible.

An additional type of reform North Carolina adopted during the period of interest consisted of increasing the accountability of students. Beginning in 2001, fifth grade students were required to satisfy a certain threshold of performance in order to be promoted to the sixth grade, and starting in 2002, grade three students were subjected to a similar requirement.[57] For this type of reform to bias my results, it would need to affect students differentially by school configuration.[58] Finally, the federal No Child Left Behind Act was implemented in North Carolina in 2003 (the 2002-03 school year). It is not obvious why this reform would lead to the patterns in the data that have been uncovered. Nevertheless, I must admit to this being a possibility, however remote.

To determine whether alternative reforms are driving my results, I carry out a falsification exercise where I counterfactually assume that the accountability reform began in a year other than 1997. The results of this exercise are reported in Table 5. Strikingly, 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. The counterfactual point estimates in 1998 are smaller than in 1997, while the estimates in 2001 and 2003 are substantially and significantly smaller than the 1997 ones. Therefore, the dynamic effects that I have uncovered are robust to the implementation of additional policies during the period of interest.

6.4 Expanding Upon the Main Results

Given the strong support that the main estimates provide for the dynamic gaming hypothesis, it is worth exploring how they manifest in various dimensions. Given the abundance of post-reform years, one feasible and interesting exercise involves analyzing the evolution of the ratchet effect over time. Table 6 reports the difference-in-differences and triple-differences estimates for three post-reform periods of equal duration, indicating that the disparity evolves as one might expect. In particular, both estimates are smallest for the initial period (1997-1999) as principals and teachers acclimate to the new incentive environment. This is true for both the comparison between K-5 and K-8, and K-5 and K-6 schools. While there is not enough power to establish this difference statistically for the triple-differences estimate, the difference-in-differences estimate for 1997-1999 is statistically smaller than the analogous one

[57] See Cooley (2010) for a more in-depth explanation of the grade five portion of the reform.

[58] While such bias is unlikely, estimates would be biased downward if students in K-8 or K-6 schools respond more strongly to the student reform than those at K-5 schools. Conversely, an upward bias would result if the opposite were true. In either case, a stronger response might arise if a greater percentage of grade five students were marginal.

for 2000-2002.

Although the triple-differences specification and the subsample restriction to only include schools that maintain the same grade configuration combine to guard against most threats to identification, an additional (but likely secondary) source of bias may stem from differences in production that vary over time and grades through, for instance, peer effects. Older students in K-8 schools are likely to have an effect on younger students that has no analog in K-5 schools.[59] If this generally deleterious effect not only changes from the pre- to post-reform period, but also operates differentially across grades, then estimates of the ratchet effect would be biased.[60]

Here, a key piece of evidence can be invoked to discount this possibility. The education literature suggests that teachers have a greater effect on mathematics than on reading scores.[61] Therefore, if the positive difference-in-differences and triple-differences estimates reflect the presence of ratchet effects rather than peer effects, one would expect mathematics scores to account for a larger proportion of the overall effect. This is what emerges. Table 7 presents estimates using both the combined score and mathematics alone and, for both comparisons across configurations (K-5 versus K-8, and K-5 versus K-6), the effect for mathematics is greater than two-thirds of the combined effect for all estimates that are significant.

Moving beyond issues of identification, it is desirable to better understand how the effects arise. 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.[62] This is exactly what one finds when dividing the difference-in-differences and triple-differences effects according to schools with a small and large number of tested classes per grade. Table 8 shows that the point estimates are larger for schools with a small number of classes per grade in all cases and significantly so for the difference-in-differences estimates.

To gain insight into the degree to which the dynamic gaming effects are centralized (coordinated via the principal), I also analyze the estimates according to whether the principal

[59] See Cook *et al.* (2008) and Bedard and Do (2005) for a discussion of these peer effects.

[60] It is possible that older students become less of a negative influence on their younger peers over time. Jacobson (2004), for example, documents a national trend of declining soft drug use among teenagers between 1997 and 2000 that supports this idea. If this trend is applicable to junior high students in North Carolina, then the disparity due to deleterious peer effects between K-5 and K-6 or K-8 schools may diminish, resulting in downward-biased estimates. Conversely, upward bias would result if the trend were reversed.

[61] For example, see Rivkin *et al.* (2005).

[62] 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.

is new to the school or more established. If the principal really is coordinating her teachers, one would expect her to be more effective given a more complete information set about the teachers. Table 9 shows that the point estimates for principals with more than one year of school-specific experience are larger than for those who are new to their school. Although this difference cannot be established statistically, owing to the large standard errors, this evidence is at least suggestive of the role that principals play in responding to the accountability scheme.

7 Structural Estimation with Linear Technology

Beyond a simple reduced-form analysis of ratchet effects, there are advantages to estimating the structural parameters of the model directly. Doing so provides a more complete understanding of the production process associated with learning and allows for illuminating counterfactual policy experiments to be conducted. A nice feature of the model specification is that the robust reduced-form results can be transformed to directly yield structural parameter estimates. In this section, I describe this transformation process for the model with fully persistent educational inputs and with partially transitory teacher inputs. I then report the structural estimates that arise using these techniques.

7.1 Structural Strategy

Abstracting away from the nonlinear scheme and technology, the structural parameters of the model can be readily expressed in terms of the difference-in-differences estimates, maintaining the benefits of the reduced-form identification strategy.[63] Using $\Phi_{cc'g} = \Delta\Delta y_{cc'g} - \Delta\Delta X'_{cc'}\beta$ and assuming that the difference-in-differences of the score, adjusted for observable characteristics, provides an unbiased measure of the distortion ($\Delta\Delta\nu_{cc'g} = 0$), equation (**??**) becomes

$$\Phi_{cc'g} = \sum_{i=0}^{g} \gamma^{g-i} \Delta\Delta e_{cc'i} \,. \tag{11}$$

Exploiting the fact that $\Phi_{cc'g}$ and $\Phi_{cc'g-1}$ are measured for $g > 1$, equation (11) can be re-expressed as

$$\Phi_{cc'g} = \gamma \Phi_{cc'g-1} + \Delta\Delta e_{cc'g} \,. \tag{12}$$

Consider the case where c and c' represents the K-5 and K-8 (or K-6) configuration, respec-

[63] Under a nonlinear scheme, there exist period-specific idiosyncratic interaction effects, which are not identified due to insufficient variation. Even estimating the average nonlinear effect for each configuration is problematic without further assumptions. Although potentially interesting, such estimates are unlikely to be of first-order importance, given the aggregated level at which the analysis occurs.

tively, so that $G_c = G = 3$. From the model, the first-order conditions for the simplifying linear scheme and production technology imply that

$$\delta\gamma\Delta\Delta e_{K5,K8,G} = \Delta\Delta e_{K5,K8,G-1} = \delta^2\gamma B(\alpha-\gamma)(1+\delta\gamma+\delta^2\gamma^2)\,, \tag{13}$$

and

$$\delta\gamma\Delta\Delta e_{K5,K6,G} = \Delta\Delta e_{K5,K6,G-1} = \delta^2\gamma B(\alpha-\gamma)\,.^{64} \tag{14}$$

Equation (12) for $g = G$ and $g = G-1$, equation (13) or (14), and the difference-in-differences estimates $\{\Phi_{K5,K8/K6,g}\}_{g=1}^{3}$ combine to produce two equations with the two unknowns γ and B, assuming α and δ are given.[65] Therefore, the structural parameters are identified from variation in scores across grades and school configurations.

The identification of structural parameters from difference-in-differences estimates extends to the case where teacher inputs and the shock are transitory with persistence $\omega\gamma_g < \gamma_g$. However, an additional identifying assumption must be made to estimate the extra parameter ω, which is that grade-specific observable student characteristics are informative as to the growth parameters γ_g. In particular, one can construct a grade-specific index ψ_g based on the observables, such that $\psi_g \equiv \bar{X}_g'\beta_g$, where $\bar{X}_g$ is a vector of average characteristics by grade. Assuming that these indices adhere to the production technology given by equation (4) of the model, the ratio of consecutive indices yields an estimate of the respective growth parameter. That is, $\gamma_g = \frac{\psi_g}{\psi_{g-1}}$.

Under the same assumption used for identification under the basic linear model with full persistence (i.e. that $\Delta\Delta\nu_{cc'g} = 0$), equation (4) can be expressed in the following difference-in-differences form:

$$\Phi_{cc'g} = \gamma_g\Phi_{cc'g-1} + \gamma_g(\omega-1)\Delta\Delta e_{cc'g-1} + \Delta\Delta e_{cc'g}\,. \tag{15}$$

Recall from the model that the first-order conditions for a school of type c are given by

$$\begin{aligned}
e_{G_c} &= B \\
e_{G_c-1} &= B\big[1+\delta(\omega\gamma_{G_c}-\alpha_{G_c})\big] \\
e_{g=G_c-\kappa} &= B\big[1+\delta(\omega\gamma_{g+1}-\alpha_{g+1}) + \delta\omega\textstyle\sum_{i=1}^{G_c-g-1}\delta^i(\gamma_{g+1+i}-\alpha_{g+1+i})\prod_{j=1}^{i}\gamma_{g+j}\big]
\end{aligned}$$

for $\kappa \geq 2$. If c and c' represents the K-5 and K-8 (or K-6) configuration, respectively, the

[64] An important assumption for these expressions to hold is that the pre-reform effort for each type of school is identical. There is not enough variation to identify separate pre-reform levels.

[65] The parameter B should be interpreted as the average myopic effect of the reform across configurations.

corresponding difference-in-differences expressions are

$$\begin{aligned}
\Delta\Delta e_{K5,K8,3} &= \delta B[\alpha_4 - \omega\gamma_4 + \delta\omega\gamma_4(\alpha_5 - \gamma_5) + \delta^2\omega\gamma_4\gamma_5(\alpha_6 - \gamma_6)] \\
\Delta\Delta e_{K5,K8,2} &= \delta^2 B\omega\gamma_3[\alpha_4 - \gamma_4 + \delta\gamma_4(\alpha_5 - \gamma_5) + \delta^2\gamma_4\gamma_5(\alpha_6 - \gamma_6)] \\
\Delta\Delta e_{K5,K8,1} &= \delta^3 B\omega\gamma_2\gamma_3[\alpha_4 - \gamma_4 + \delta\gamma_4(\alpha_5 - \gamma_5) + \delta^2\gamma_4\gamma_5(\alpha_6 - \gamma_6)]
\end{aligned}$$

and

$$\begin{aligned}
\Delta\Delta e_{K5,K6,3} &= \delta B(\alpha_4 - \omega\gamma_4) \\
\Delta\Delta e_{K5,K6,2} &= \delta^2 B\omega\gamma_3(\alpha_4 - \gamma_4) \quad . \\
\Delta\Delta e_{K5,K6,1} &= \delta^3 B\omega\gamma_2\gamma_3(\alpha_4 - \gamma_4)
\end{aligned}$$

As before, I assume that the pre-reform effort for each type of school is identical. Combining either set of conditions with equation (15) for $g = G$ and $g = G - 1$, the estimates $\{\Phi_{K5,K8/K6,g}\}_{g=1}^{3}$ and $\gamma_g = \frac{\psi_g}{\psi_{g-1}}$ for $g \in \mathcal{G}_{c'=K8/K6}$, there are two equations containing the two unknowns ω and B, assuming α and δ are given. Therefore, in this more general case, the structural parameters are also identified from variation in scores across grades and school configurations.

7.2 Linear Estimates

I first present structural parameter estimates for the model with linear technology and persistent inputs, by transforming the difference-in-differences estimates as per the previously outlined strategy. I estimate the model using the actual value of the target $\alpha = 0.924$,[66] and assume an inter-temporal depreciation parameter δ of 0.9.[67] Table 10 presents the structural estimates for each configuration comparison using specification (3) with full controls. For the comparison between K-5 and K-6 schools, the growth parameter γ is estimated to be 0.54, the myopic parameter B is estimated to be 5.30, and both parameters are highly significant.

Given the value of the target coefficient α, the estimate for the growth parameter may seem low. However, one must remember that this basic estimated model does not allow for separate growth rates in parental and teacher inputs. It is generally understood that the former type of inputs grow at a higher rate than the latter type, with various studies placing an upper bound on the persistence of teacher effects at 50 percent per year.[68] Therefore, the estimate

[66] Given the subject- and grade-specific coefficients outlined in Section 4, the equivalent expected growth coefficient for the combined reading and mathematics score is 0.88. The average of the expected and high (10% higher) growth coefficient is then $\alpha = 1.05 \times 0.88 = 0.924$.

[67] In practice, the estimates are fairly insensitive to the choice of δ, which is not separately identified in the model.

[68] See Jacob *et al.* (2008), Kane and Staiger (2008), and Rothstein (2010), for example.

for γ should be interpreted as a weighted average of the growth rates for each input type. The parameters are not as precisely estimated for the comparison between K-5 and K-8 schools. Given that structural identification depends on the precision of the underlying difference-in-differences estimates for each of the shared grades, such imprecision is not surprising. While estimates are highly significant for all grades when comparing K-5 and K-6 schools, the same cannot be said for grades three and four when comparing K-5 and K-8 schools.

With an additional identifying assumption, the structural estimation strategy can be extended to allow for differential growth rates between teacher and non-teacher inputs. Specifically, as shown in the econometric framework, the persistence parameter ω is identified if the overall growth parameters are determined from variation in observable student characteristics. Given the actual target $\alpha = 0.924$ and using $\delta = 0.9$, ω and B are recovered by calculating each γ_g from the observable indices ψ_g and transforming the estimates $\{\Phi_{K5,K8/K6,g}\}_{g=1}^{3}$ according to equation (15) and the relevant difference-in-differences first-order conditions.

Table 11 presents the structural estimates of the transitory model, using a full set of controls (specification (3)). The additional parental and student ethnicity controls ensure that the growth parameters are estimated precisely, although not necessarily in an unbiased way. The extent of bias will be determined directly by how well the growth in observables approximates the growth in all variables, including unobservables. Although this approximation cannot be verified, it is still informative to estimate the model in this way, as it allows an additional degree of freedom with which to identify the transitory parameter ω. The average of the growth rates for grades four, five and six is 0.853, which is substantially higher than the 0.56 estimate that arises from the more restrictive persistent model and is more in line with the actual accountability target of 0.924. The estimate for B is 5.87, but insignificant. This is of the same magnitude found in the fully persistent analysis. Interestingly, the estimate for ω from the structural analysis is 0.85 and is significant at the 10 percent level. This means that the effect of teacher effort on student achievement diminishes at a 15 percent greater rate than other inputs. This is in keeping with the previously mentioned literature on teacher effects.

8 Structural Estimation with Nonlinear Technology

Although standard in the education literature, a linear production technology may fail to capture important complementarities in production. I consider one such interaction between teacher effort and student ability, as reflected by the student prior score, which is conveniently represented by the production function in equation (5). Doing so allows me to test this

model against the simpler linear one, by ascertaining whether the latter can be statistically rejected.[69] If so, it is informative to establish whether the underlying inputs are complements or substitutes in production. Beyond gaining insight into the learning process, this exercise is also important for determining how the interaction affects the magnitude of the ratchet effect when compared to the linear case. In what follows, after discussing the structural estimation method and identification strategy, I present the results of the analysis.

8.1 Structural Strategy

The inherent nonlinearity of a specification with interactions between teacher effort and student ability demands a more sophisticated estimation technique than a simple transformation of difference-in-differences estimates. To that end, I employ a maximum-likelihood approach with embedded fixed effects to control for unobserved differences across grade horizons. The estimation problem is to select the parameter values that maximize the log-likelihood function

$$\mathcal{L}(\gamma, \theta, B, \sigma^2) = \sum_{t=1}^{T} \sum_{s=1}^{S} \sum_{g \in G_c} ln(\varphi(u_{scgt}; \gamma, \theta, B, \sigma^2)\,, \tag{16}$$

where $u_{scgt} = y_{scgt} - \gamma y_{scg-1t-1} - \theta e_{scgt} y_{scg-1t-1} - e_{scgt} - a_{scgt}$ from equation (5), $\varphi(\cdot)$ is the density function of the shock u that is normally distributed with mean zero and variance σ^2, and e_{scG_ct} and e_{scG_c-1t} are given by equations (6) and (7).[70,71]

The fixed effects are designed to account for differing unobserved teacher ability a_{scgt}. Due to the incidental parameters problem, it is not possible to identify each idiosyncratic effect. Instead, I include grade horizon fixed effects, which account for differences at the horizon level that are not attributable to differential effort, such as teacher ability.[72] This is done over all available time periods so that fixed effects are identified from the pre-reform period when effort due to the reform is assumed to be zero. While the approach does not account for common trends over time, as in the difference-in-differences linear approach, its greater flexibility allows for the identification of the additional interaction parameter θ. Having controlled for fixed effects, there are essentially three types of school in the likelihood function over which effort is expected to vary: schools serving the final grade G, the second-from-last

[69] A useful feature of the nonlinear specification is that the linear model is a special case and can be easily recovered by setting $\theta = 0$.

[70] As is apparent from equation (16), shocks are assumed to be serially uncorrelated over time and grades. While the former is unlikely to be an issue, the latter may be. I plan to address this in future work.

[71] The relevant equation for e_{scG_c-2t} is omitted here due to complexity, but the quantity is simulated in the maximum-likelihood routine.

[72] For the purposes of the counterfactual policy experiments that follow, including fixed effects at the configuration-grade level, instead of the horizon level, produces quantitatively similar results.

grade $G-1$ and the third-from-last grade $G-2$. Conditional on the score for the prior grade $y_{scg-1t-1}$, the distinctly different ratcheting behavior for each of these horizons, as captured by the associated first-order conditions, is what identifies the three structural parameters of interest γ, θ and B. Confidence intervals for each estimate are then bootstrapped using repeated samples from the error structure, as implied by the model and point estimates.

8.2 Nonlinear Estimates

As discussed, I estimate the model using maximum-likelihood estimation, embedding fixed effects to control for unobserved differences across grade horizons. I utilize the quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) gradient method to solve the optimization problem. Taking the resulting estimates for the growth parameter γ, the interaction parameter θ and the myopic parameter B as given, I then infer the underlying error structure. Using 4000 draws with replacement, this is used to construct a bootstrap distribution for each parameter, by obtaining parameter estimates that maximize the likelihood function for each draw. This allows for corresponding percentile confidence intervals to be computed.

The results are shown in Table 12 for $\alpha = 0.924$ and $\delta = 0.9$.[73] Confidence intervals at the 90 and 95 percent level are reported. The estimate for γ is 0.869, while B is estimated to be 1.18. Both are significant at the 95 percent level and in line with the magnitudes already established. Additionally, the fact that γ is estimated to be less than the target α supports the idea that effort is increasing in the grade. More interestingly, the parameter θ is estimated to be 0.0019, rejecting the more restrictive linear technology hypothesis at the 95 percent level. This positive value suggests that teacher effort and student ability are complements in production, which is a novel finding in the education literature.

To compare the distortion under the linear and nonlinear models, I re-estimate the model with the restriction $\theta = 0$. The resulting linear estimates are 0.875 and 2.90 for γ and B, respectively. Proposition 8 implies that the distortion should be attenuated for $\theta > 0$ if γ is unchanged across the comparison. However, γ is estimated to be lower under the nonlinear specification, which should lead to an increased distortion in isolation. Despite this, the complementarity is sufficiently strong to create an overall distortion for a given prior score $y_{scg-1t-1}$ that is smaller under the less restrictive nonlinear specification.[74] Using the structural estimates, the average distortion for grade three and four in the nonlinear case is 0.079 and

[73] The choice of δ does not substantively affect the estimates.

[74] The distortion in effort is approximately equal to $\delta[\gamma - \alpha + 2B\theta(1 + \gamma\theta y_{scG_c-2t-1})]$ for grade $G_c - 1$ and is more involved for lower grades.

0.044, respectively, while the analogous quantities for the more restrictive linear case are 0.080 and 0.045. Thus, the distortion between grades is slightly overestimated by assuming the technology is linear.

9 Policy Experiments

The structural estimates provide useful insight into the technology that underlies the learning process (not least, the potential nonlinearity just discussed). An additional reason for going beyond a reduced-form analysis of ratchet effects is to carry out informative policy experiments.

The first experiment (which I refer to as Simulation A) involves exploring a counterfactual world in which the reform was never enacted. This sheds light on the full effect of the reform, accounting for the cumulative nature of educational inputs in the production process. Counterfactually setting the parameter B equal to zero, effort from the reform becomes zero in every grade. The corresponding results are presented in the Simulation A panel of Table 13. Using the general nonlinear structural estimates, $\gamma = 0.869$, $B = 1.18$ and $\theta = 0.0019$, the resulting cumulative grade five score at the average K-5 school is 44.3% of a standard deviation lower than the actual level that is observed. Thus, in keeping with the descriptive evidence, the counterfactual evidence indicates that the reform had a substantial effect on student achievement.

The second experiment (Simulation B) uses the theoretical prediction that the ratchet effect in grade $G_c - 1$ is eliminated for the average student by choosing the target $\alpha = \gamma + 2B\theta(1 + \gamma\theta\bar{y}_{cG_c-2t-1})$, where $\bar{y}_{cG_c-2t-1}$ is the average prior score for school configuration c.[75] On this basis, I can quantify the cumulative effect of the dynamic distortions on the grade five score. By eliminating distortions at the average K-5 school, the effort level is unchanged in grade five (which was undistorted to begin with) and rises in grades three and four. These changes for lower grades are shown in the bottom sub-panel of the Simulation B panel of Table 13. In accordance with earlier reduced-form evidence, the effect of the distortion is greater in grade three than in grade four. The increases in early grade effort have a compounding effect on the grade five score due to the role of the production technology, which is presented in the top sub-panel of the Simulation B panel of Table 13. The cumulative effect of eliminating ratcheting behavior at the average K-5 school amounts to a 1.7% of a standard deviation increase in the grade five score. However, such a scheme is about 37% more

[75]Ratchet effects in lower grades are eliminated in an analogous way. The associated counterfactual targets are more involved and are omitted here, but they are implemented analytically in the policy experiment.

costly to implement, as the target α is lowered in this instance to thwart dynamic gaming, making it easier to satisfy.

There are alternative ways to formulate the relevant nonlinearities. The chosen specification proves to be analytically tractable for the purposes of comparing the counterfactuals in simulations A and B with their linear counterparts, which allows one to gauge whether the linear specification provides a close approximation. These comparisons are found in both panels of Table 13. Setting $\theta = 0$ under the general model, the linear parameter estimates are $\gamma = 0.875$ and $B = 2.90$. Using these values, a world without the reform would see the cumulative grade five score fall by 45.3% of a standard deviation, which overstates the more flexible nonlinear result (0.44 standard deviations) by about 2.3%. This can be explained by the biased estimates under the restricted linear case and the production technology specified in equation (5),[76] taking advantage of the structural formulation. The counterfactual score decomposition for Simulation A reveals that the upward-biased estimates of γ and B result in an understated teacher ability fixed effect and an overstated educational capital component $\gamma\bar{y}_4$ under the linear restriction. Overall, the former effect dominates the latter one, leading to a lower counterfactual score and thus an overstatement of the reform's effect relative to the nonlinear specification.

Counterfactually eliminating the distortion in Simulation B, I find that the cumulative increase in the grade five score is 1.8% of a standard deviation. Comparing this result to the more flexible nonlinear result, the linear simplification overstates the cumulative effect of eliminating the distortion by 5.8%. The structure of the model makes the origin of this overstatement evident from the corresponding counterfactual score decomposition. As with the Simulation A decomposition, the upward-biased growth rate parameter γ results in a larger educational capital component $\gamma\bar{y}_4$ and the teacher fixed effect is unchanged. However, the effective effort term ($\tilde{e}_5 = e_5(1 + \theta\bar{y}_4)$) is no longer equal to zero. Instead, it reflects the dynamically undistorted effect of effort, which is larger for the nonlinear case. Although the grade-by-grade distortion is greater under the linear restriction, as reflected by the higher cost associated with eliminating the distortions (37.3% compared to 36.9% under the nonlinear benchmark), the effect of the distortions is larger without the restriction, mainly owing to the additional contribution that the interaction term brings to effective effort. Despite the smaller effective effort under the linear case, the educational capital term is sufficiently large so that the score gain from eliminating the distortion is overstated compared to the nonlinear specification.

[76]The general technology is $y_{scgt} = \gamma y_{scg-1t-1} + a_{scgt} + e_{scgt}(1 + \theta y_{scg-1t-1}) + u_{scgt}$.

This structural approach highlights the potential for model misspecification to affect educational policy evaluation in unexpected ways. It allows for multiple interconnected factors to be disentangled, revealing that a predicted overstatement of a policy change influenced by reduced-form intuition remains an overstatement in practice. The nontrivial disparities outlined in this analysis suggest caution is warranted when adopting a linear technology approximation in other educational contexts.

10 Conclusion

Value-added incentive schemes have been used with increasing frequency under a multitude of accountability reforms enacted over the past two decades. The chief benefit of the performance targets that are central to these schemes is that they adjust for unobserved heterogeneity in scholastic inputs. Yet a rich dynamic incentive theory literature predicts that the inherent inter-temporal dependence of these targets should engender dynamic gaming of effort, known as the ratchet effect. Given the substantial stakes associated with accountability schemes, it is important for policymakers to understand whether ratchet effects arise in practice and if so, how much they distort outcomes. No analyses have explored this issue. Even outside the educational literature, very few studies have attempted to reconcile the relevant theory with empirical evidence.

A primary reason for this state of affairs is that existing theoretical formulations 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 first extend the theoretical literature to include ratchet effects with finite horizons, intentionally capturing salient features of value-added accountability reforms. This exercise produces a viable research design, where ratchet effects are identified from variation in the horizon schools face, as captured by the school grade span. Using a triple-differences strategy, I find substantial evidence of such effects, with distortions ranging between 3.9% and 5.9% of a standard deviation in the grade five score. These new dynamic results expand on the established (static) educational gaming literature.

Going beyond the reduced-form analysis, I also structurally estimate the model. Doing so provides insight into the technology that underlies the learning process and makes informative counterfactual policy experiments possible, based on a more general education technology. In one experiment, I determine that the grade five score would have been 44.3% of a standard deviation lower if the reform had not been implemented. A second experiment uses a key

finding that emerges from the theory, revealing how to eliminate the dynamic distortion while maintaining the desirable aspects of the reform. Applying that theoretical result, I find that the grade five score would be 1.7% of a standard deviation higher in the absence of ratchet effects, but would also be about 37% more expensive to implement, making it a relatively undesirable remedy for policymakers.

The results of this analysis have implications for a broader research agenda that focuses on the production technology of learning in schools. Throughout the course of uncovering whether educational accountability has dynamic effects on achievement, several tantalizing aspects of the learning process have come to light, including input-specific effort persistence and complementarities between inputs. These issues are likely to have a significant effect on the design of efficient education policies. Moreover, they present a rare opportunity to infer idiosyncratic effort — typically a challenging task — using a procedure guided by the dynamic framework that I have developed. Inference of such a quantity is valuable, since it is thought to be a key ingredient in the educational development of children. Given the importance of the aforementioned issues, I intend to examine them in my future work.

References

Ahn, Tom (2009), "The Missing Link: Estimating the Impact of Incentives on Effort and Effort on Production Using Teacher Accountability Legislation," working paper, *http://sites.google.com/site/tomahnjobmarket/test/TeacherEffort-working.pdf*.

Allen, Douglas W. and Dean Lueck (1999), "Searching for Ratchet Effects in Agricultural Contracts," *Journal of Agricultural and Resource Economics*, 24(2): 536-552.

Alspaugh, John W. (2001), "Achievement Loss Associated with the Transition to Middle School and High School," *Journal of Educational Research*, 92: 20-25.

Barlevy, Gadi and Derek Neal (2011), "Pay for Percentile," NBER Working Paper No. 17194, July.

Baron, David P. and David Besanko (1987), "Commitment and Fairness in a Dynamic Regulatory Relationship," *Review of Economic Studies*, 54(3): 413-436.

Bedard, Kelly and Chau Do (2005), "Are Middle Schools More Effective? The Impact of School Structure on Student Outcomes," *Journal of Human Resources*, 40(3): 660-682.

Bifulco, Robert and Helen F. Ladd (2006), "The Impacts of Charter Schools on Student Achievement: Evidence from North Carolina," *Education Finance and Policy*, 1(1): 50-90.

Carnoy, Martin and Susanna Loeb (2002), "Does External Accountability Affect Student Outcomes? A Cross-State Analysis," *Educational Evaluation and Policy Analysis*, Winter, 24(4): 305-331.

Charness, Gary, Peter Kuhn and Marie Claire Villeval (2010), "Competition and the Ratchet Effect," NBER Working Paper No. 16325, September.

Cooley, Jane (2010), "Desegregation and the Achievement Gap: Do Diverse Peers Help?" working paper, *http://www.ssc.wisc.edu/~jcooley/CooleyDeseg.pdf*.

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(4): 781-801.

Cook, Phillip J., Robert MacCoun, Clara Muschkin and Jacob Vigdor (2008), "The Negative Impacts of Starting Middle School in Sixth Grade," *Journal of Policy Analysis and Management*, 27(1): 104-121.

Cullen, Julie B. and Randall Reback (2006), "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System," NBER Working Paper No. 12286, June.

Efron, B. and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.

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," Doctoral Dissertation, N.C. State University, *http://www.lib.ncsu.edu/resolver/1840.16/5963*.

Figlio, David N. and Lawrence W. Kenny (2007), "Individual Teacher Incentives and Student Performance," *Journal of Public Economics*, 91(5-6): 901-914.

Freixas, Xavier, Roger Guesnerie and Jean Tirole (1985), "Planning Under Incomplete Information and the Ratchet Effect," *Review of Economic Studies*, 52(2): 173-191.

Gibbons, Robert (1987), "Piece-Rate Incentive Schemes," *Journal of Labor Economics*, 5(4): 413-429.

Gibbons, Robert (1996), "Incentives and Careers in Organizations," NBER Working Paper No. 5705, August.

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(9-10): 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(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(1): 43-59.

Holmstrom, Bengt (1982), "Design of Incentive Schemes and the New Soviet Incentive Model," *European Economic Review*, 17: 127-148.

Jacob, Brian A., Lars Lefgren and David Sims (2008), "The Persistence of Teacher-Induced Learning Gains," NBER Working Paper No. 14065, June.

Jacob, Brian A. and Steven Levitt (2003), "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics*, 118(3): 843-877.

Jacobson, Mireille (2004), "Baby Booms and Drug Busts: Trends in Youth Drug Use in the United States," *Quarterly Journal of Economics*, 119(4): 1481-1512.

Juvonen, Jaana, Vi-Nhuan Le, Tessa Kaganoff, Catherine Augustine and Jouay 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," NBER Working Paper No. 8156, March.

Kane, Thomas J. and Douglas O. Staiger (2002), "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives*, 16(4): 91-114.

Kane, Thomas J. and Douglas O. Staiger (2008), "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," NBER Working Paper No. 14607, December.

Kanemoto, Yoshitsugu and W. Bentley MacLeod (1992), "The Ratchet Effect and the Market for Secondhand Workers," *Journal of Labor Economics*, 10(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(4): 347-367.

Ladd, Helen F. (2001), "School-based Educational Accountability Systems: The Promise and the Pitfalls," *National Tax Journal*, 54(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(4): 494-529.

Laffont, Jean-Jacques and Jean Tirole (1988), "The Dynamics of Incentive Contracts," *Econometrica*, 56(5): 1153-1175.

Lavy, Victor (2002), "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement," *Journal of Political Economy*, 110(6): 1286-1317.

Lavy, Victor (2009), "Performance Pay and Teachers' Effort, Productivity and Grading Ethics," *American Economic Review*, 99(5): 1979-2011.

Lazear, Edward P. (1986), "Salaries and Piece Rates," *Journal of Business*, 59(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(1), Value-Added Assessment Special Issue, Spring: 67-101.

Muralidharan, Karthik and Venkatesh Sundararaman (2009), "Teacher Performance Pay: Experimental Evidence from India," NBER Working Paper No. 15323, September.

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

Parent, Daniel (1999), "Methods of Pay and Earnings: A Longitudinal Analysis," *Industrial and Labor Relations Review*, 53(1): 71-86.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain (2005), "Teachers, Schools, and Academic Achievement," *Econometrica*, 73(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(11-12): 1051-1061.

Rothstein, Jesse (2010), "Teacher Quality in Educational Production: Tracking, Decay and Student Achievement," *Quarterly Journal of Economics*, 125(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(485): F3-F33.

Todd, Petra E. and Kenneth I. Wolpin (2007), "The Production of Cognitive Achievement in Children: Home, School, and Racial Test Score Gaps," *Journal of Human Capital*, 1(1): 91-136.

Weitzman, Martin L. (1980), "The Ratchet Principle and Performance Incentives," *Bell Journal of Economics*, 11(1): 302-308.

Table 1: Descriptive Statistics

Variable	**Mean**	**St. Dev.**	**Min**	**Max**
Combined Math and Reading Score:				
Grade 3	291.5	18.8	215	345
Grade 4	303.3	18.3	229	357
Grade 5	314.7	16.9	234	368
Grade 6	322.2	18.8	254	376
Grade 7	331.2	18.0	261	386
Grade 8	337.6	18.4	271	395
Student - Parental Education:				
No High School	0.10	0.30	0	1
High School Graduate	0.42	0.49	0	1
Trade School	0.09	0.28	0	1
Community College	0.11	0.32	0	1
4-Year College	0.22	0.42	0	1
Graduate Degree	0.06	0.23	0	1
Student - Ethnicity:				
White	0.64	0.48	0	1
Black	0.28	0.45	0	1
Other	0.08	0.27	0	1
Student - Exceptionality:				
Learning Impairment	0.12	0.32	0	1
No Special Label	0.74	0.44	0	1
Gifted	0.14	0.35	0	1
Teacher:				
Experience*	12.6	6.0	0	42
License Test Score*	0.00	0.55	-3.42	3.20
School - Locale:				
Large City	0.06	0.23	0	1
Mid-Size City	0.21	0.41	0	1
Urban Fringe of Large City	0.05	0.21	0	1
Urban Fringe of Mid-Size City	0.13	0.34	0	1
Small Town & Rural	0.55	0.50	0	1
School - Other:				
% Free Lunch Eligible	0.38	0.20	0	0.99
School-Specific Principal Tenure*	3.3	2.2	1	11
No. of Classes Per Grade (Gr. 3-5)*	3.5	1.4	1	13

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). Student and school location categories are both mutually exclusive and exhaustive.

Table 2: Reduced-Form 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)
School - % Free Lunch Eligible		-5.24**	-4.83**
		(0.35)	(0.36)
Teacher - License 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	6130308	5318520	4499997

Note: This table defines three specifications according to the components included in the control vector of the main estimating equation (equation (8)) 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 3. The analysis is conducted for the years 1994 through 2005, with the number of observations declining as regressors with missing values are added. For instance, specification (3) includes a control for teacher ability that is not observed in 1994. Thus, specifications (2) and (3) are presented to show that the results are robust when choosing between including a teacher control and an additional pre-reform year.

Standard errors adjusted for clustering at school level are reported in parenthesis.

Significance levels: ∗∗ denotes 1%; ∗ denotes 5%

Table 3: Reduced-Form Results

	c = K5 vs. c' = K8			c = K5 vs. c' = K6		
Specification:	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**
Grade 5 Diff-in-Diff						
$\Phi_{c,5,post-pre}$	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 Diff-in-Diff						
$\Phi_{c,4,post-pre}$	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 Diff-in-Diff						
$\Phi_{c,3,post-pre}$	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 2 and according to grade, 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 school level are reported in parenthesis.

Significance levels: ** denotes 1%; * denotes 5%; † denotes 10%

Table 4: 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 Configuration 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 to those for the subsample of schools that maintain a stable grade configuration over the period of interest. As before, the estimates are constructed from joint F-tests of the interaction dummies included in the regression.

Standard errors adjusted for clustering at school level are reported in parenthesis.

Significance levels: ** denotes 1%; * denotes 5%; † denotes 10%

Table 5: 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 school level are reported in parenthesis.

Significance levels: ** denotes 1%; * denotes 5%; † denotes 10%

Table 6: 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 school level are reported in parenthesis.

Significance levels: $**$ denotes 1%; $*$ denotes 5%; † denotes 10%

Table 7: 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 school level are reported in parenthesis.

Significance levels: $**$ denotes 1%; $*$ denotes 5%; † denotes 10%

Table 8: Coordination/Free-Riding Effects

	c = K5 vs. c' = K8			c = K5 vs. c' = K6		
# of classes / 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 school level are reported in parenthesis.

Significance levels: ** denotes 1%; * denotes 5%; † denotes 10%

Table 9: School-Specific Principal Experience

	c = K5 vs. c' = K8			c = K5 vs. c' = K6		
# 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 school level are reported in parenthesis.

Significance levels: ** denotes 1%; * denotes 5%; † denotes 10%

Table 10: Structural Results - Fully Persistent Linear Technology

	K-5 vs. K-8	**K-5 vs. K-6**
γ	0.39	0.54**
	(0.42)	(0.16)
B	1.80**	5.30**
	(0.44)	(1.22)

Note: This table presents structural parameter estimates for the linear technology model with fully persistent inputs. The parameters are estimated from a transformation of the reduced-form coefficients with full controls (specification (3)), using $\delta = 0.9$ and $\alpha = 0.924$.

Standard errors adjusted for clustering at school level are reported in parenthesis.

Significance levels: ** denotes 1%

Table 11: Structural Results - Linear Tech. with Transitory Effort (K-5 vs. K-6)

γ_4	0.84**
	(0.18)
γ_3	0.87**
	(0.06)
γ_2	0.85**
	(0.06)
ω	0.85†
	(0.45)
B	5.87
	(16.33)

Note: This table presents structural parameter estimates for the linear technology model with partially transitory teacher inputs. The parameters are estimated from a transformation of the reduced-form coefficients with full controls (specification (3)) and the subsample restriction that schools do not switch grade span, using $\delta = 0.9$ and $\alpha = 0.924$. Standard errors adjusted for clustering at the school level are reported in parenthesis.
Sig. levels: ** denotes 1%; † denotes 10%

Table 12: Structural Results - Nonlinear Technology

	2.5%	**5%**	**Pt. Est.**	**95%**	**97.5%**
γ	0.8622	0.8631	0.8685	0.8741	0.8748
θ	0.0000	0.0002	0.0019	0.0064	0.0081
B	0.2525	0.3499	1.1832	2.5166	2.8413
σ	4.2128	4.2196	4.2558	4.2907	4.2978

Note: This table presents structural parameter estimates for the model with nonlinear production technology. Parameters are estimated using maximum-likelihood estimation for $\delta = 0.9$ and $\alpha = 0.924$. Confidence bounds are obtained from the relevant percentiles of the bootstrap distribution, computed using 4000 draws from the underlying error structure.

Table 13: Counterfactual Simulations

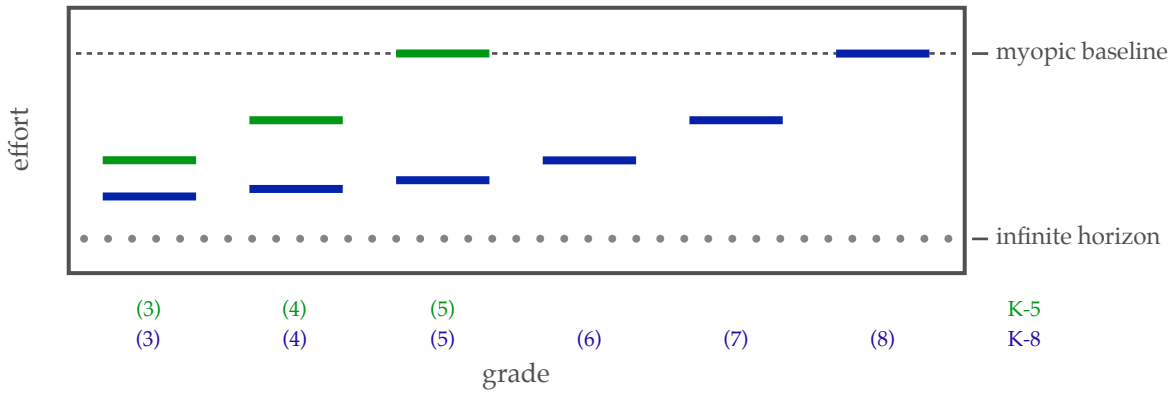
Actual Gr. 5 Score (Post-Reform):		
Mean	316.59	
Standard Deviation	**16.27**	
Simulation A: No Reform		
	Nonlinear	**Linear**
Counterfactual Score	309.39	309.22
Δ in Score†	-7.20	-7.36
Δ as % of St. Dev.	-44.3%	-45.3%
Counterfactual Score Decomposition:		
$\gamma\bar{y}_4$	260.85	262.46
Ability FE (a_5)	48.54	46.76
Effective Effort ($\tilde{e}_5$)	0.00	0.00
Counterfactual Score	309.39	309.22
Simulation B: No Distortion		
	Nonlinear	**Linear**
Counterfactual Score	316.86	316.88
Δ in Score†	0.27	0.29
Δ as % of St. Dev.	1.68%	1.78%
Counterfactual Score Decomposition:		
$\gamma\bar{y}_4$	265.37	267.22
Ability FE (a_5)	48.54	46.76
Effective Effort ($\tilde{e}_5$)	2.95	2.90
Counterfactual Score	316.86	316.88
Δ in Gr. 4 from No Distortion††	0.08	0.13
Δ in Gr. 3 from No Distortion††	0.14	0.23
% Δ in Cost of Reform	36.86%	37.32%

Note: Analysis is done on a cumulative basis for grade five scores in K-5 schools. Each decomposition is presented for the grade five score, taking the counterfactual scores in grades three and four as given. The nonlinear simulation uses estimates $\gamma = 0.869$, $B = 1.18$ and $\theta = 0.0019$, while the linear specification restricts $\theta = 0$ and utilizes the resulting estimates $\gamma = 0.875$ and $B = 2.90$. As before, the actual target is $\alpha = 0.924$ and the discounting value is $\delta = 0.9$.

†: Δ in score = counterfactual score - actual score

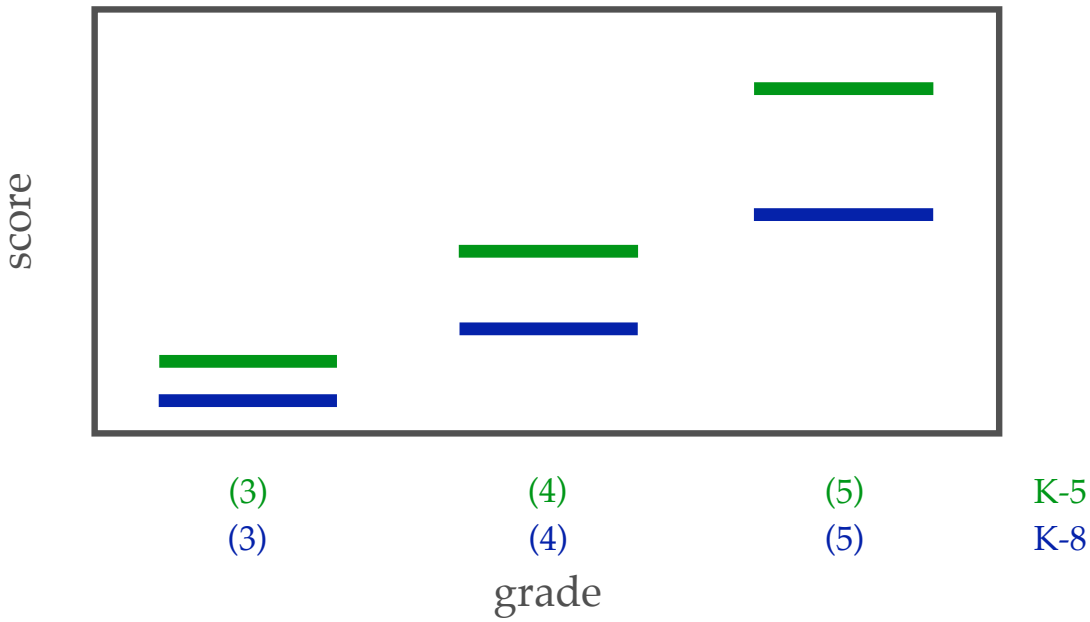
††: This is the contemporaneous effect of eliminating the distortion (approximately $B\delta(1+\theta y_{scG_c-2t-1})[\gamma - \alpha + 2B\theta(1+\gamma\theta y_{scG_c-2t-1})]$ for grade four).

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



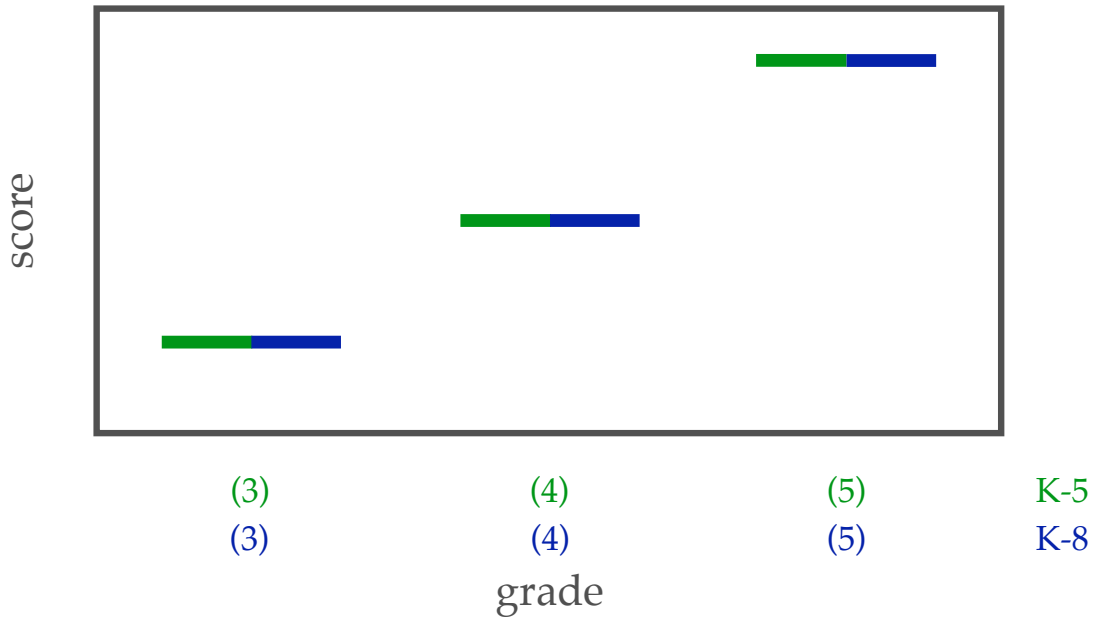
Assuming that the target coefficient exceeds the natural growth rate ($\alpha > \gamma$), this diagram contrasts the effort levels by grade for two different grade spans (as implied by the first-order conditions). This is done to illustrate how differing horizons affect the effort level for a particular 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 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.

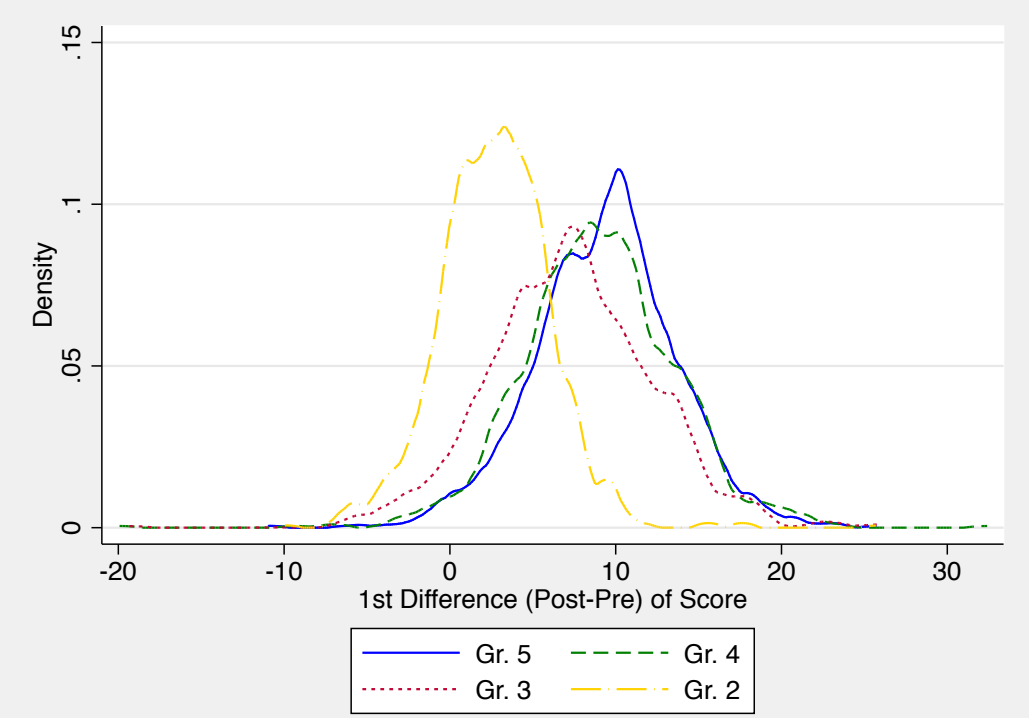
Figure 3: Eliminating Dynamic Distortions



This figure illustrates the effect of lowering the target coefficient α to be equal to the natural growth rate γ. When this is done, the scores coincide for different school grade spans with otherwise identical characteristics.

Figure 4: Density of First-Differenced Scores By Grade — All Schools

(a) Raw Score



(b) Adjusted Score

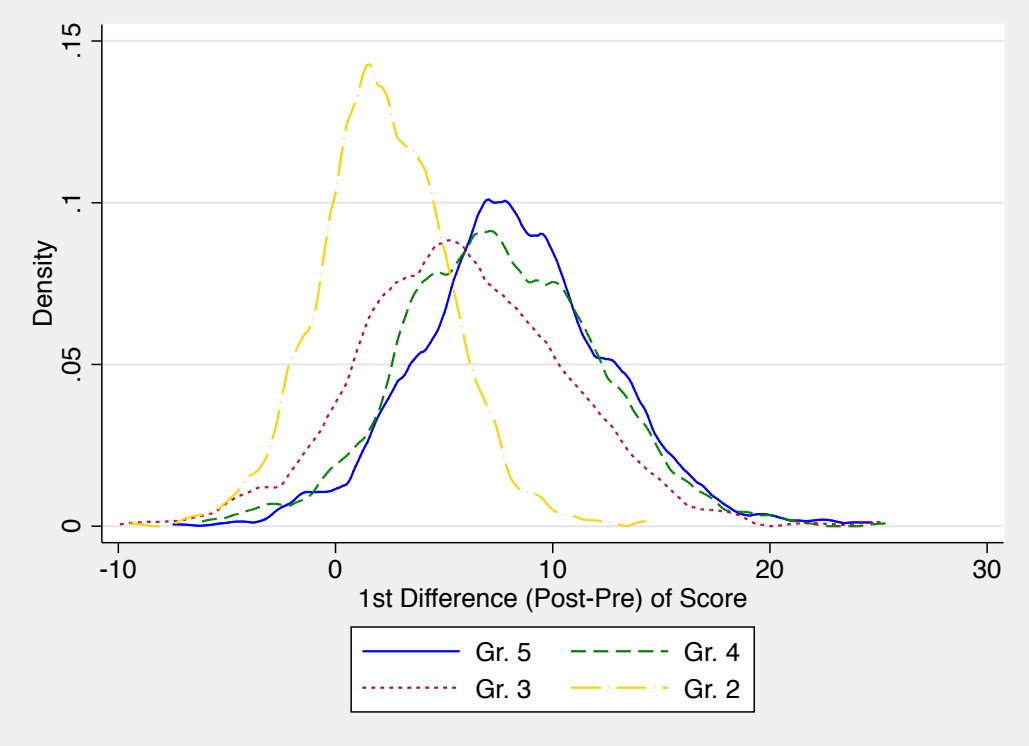
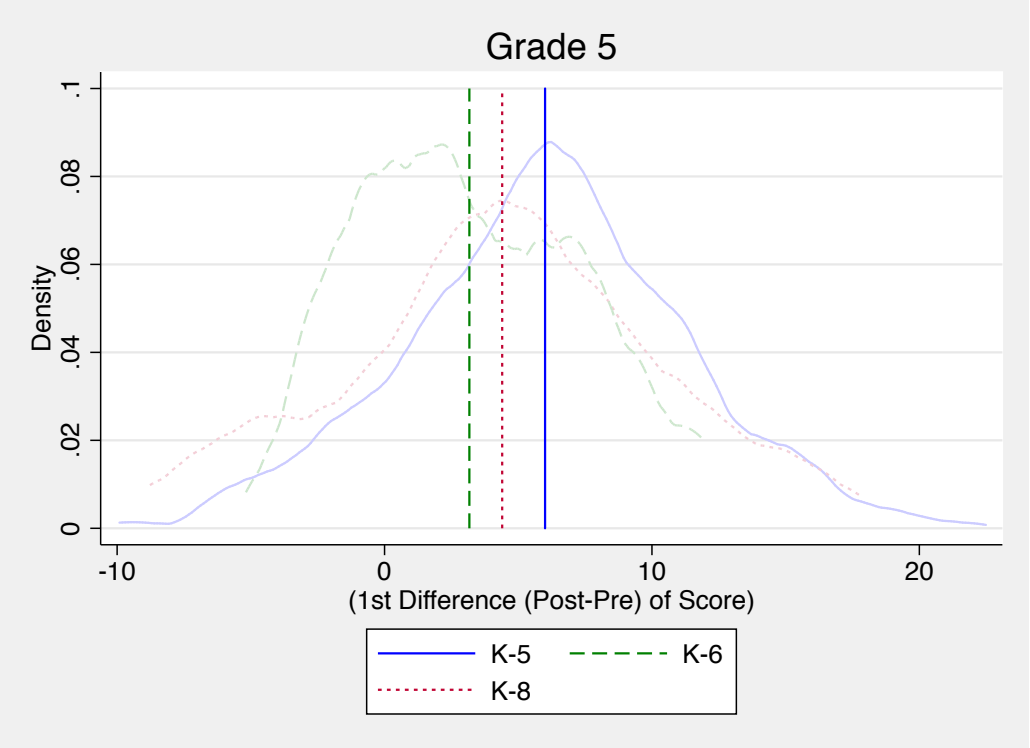


Figure 5: Grade 5 Distribution of First-Differenced Scores By Grade Span



Appendix A

A.1 Grid of Available Data

The following grid is a graphical representation of the available data by year and cohort.

Year \ Cohort	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1993	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1994	4	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-
1995	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	2
2008	-	-	-	-	-	-	-	-	-	-	8	7	6	5	4	3

For the 1995-96 school year, the data are sparse. Specifically, I only observe grade two, three and four scores for that year. The double horizontal separator following the 2004-05 school year reflects the fact that the reform was substantially altered in the following year. Although scores are comparable across the 2004-05 and 2005-06 school years (on the same developmental scale), the incentives may not be.