Submission Number: PET11-11-00042

The dynamic effects of educational accountability

Hugh Macartney
University of Toronto

Abstract

Holding educators more accountable for the academic achievement of their students has been a central feature of recent education reforms. In several prominent instances, accountability schemes have set pecuniary performance targets that condition on prior scores as a means of controlling for student heterogeneity. Yet doing so introduces a potential dynamic distortion in incentives: teachers may be less responsive to the reform today in an effort to avoid more onerous targets in future — an instance of the so-called `ratchet effect.' Given an environment where school-level targets depend on student prior scores, I show that such dynamic gaming behaviour depends crucially on variation in the horizon, with teachers distorting their effort less when their decision affects fewer future scores within the same school. Making use of rich educational panel data from North Carolina, I exploit variation in the grade span of schools to identify this effect, finding compelling evidence of dynamic distortions using a difference-in-differences approach. I then directly estimate the structural parameters of the corresponding model, allowing for complementarities in production between teacher effort and student ability. Using these estimates, the grade five score in K-5 schools would be about 1.25 standard deviations lower under a counterfactual setting without any accountability scheme and 4.6% of a standard deviation higher if ratchet effects were eliminated through target reduction.

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, Damon Clark, Stephen Coate, Elizabeth Dhuey, Weili Ding, Raquel Fernández, Amy Finkelstein, Sacha Kapoor, Steven Lehrer, Thomas Lemieux, Joshua Lewis, Enrico Moretti, Alvin Murphy, Parag Pathak, Nancy Qian, Jesse Rothstein, 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 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.
Submitted: February 17, 2011.

The Dynamic Effects of Educational Accountability *

Hugh Macartney †
Department of Economics
University of Toronto

February 17, 2011

* 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, Damon Clark, Stephen Coate, Elizabeth Dhuey, Weili Ding, Raquel Fernández, Amy Finkelstein, Sacha Kapoor, Steven Lehrer, Thomas Lemieux, Joshua Lewis, Enrico Moretti, Alvin Murphy, Parag Pathak, Nancy Qian, Jesse Rothstein, 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 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@utoronto.ca

Abstract

Holding educators more accountable for the academic achievement of their students has been a central feature of recent education reforms. In several prominent instances, accountability schemes have set pecuniary performance targets that condition on prior scores as a means of controlling for student heterogeneity. Yet doing so introduces a potential dynamic distortion in incentives: teachers may be less responsive to the reform today in an effort to avoid more onerous targets in future — an instance of the so-called 'ratchet effect.' In order to determine whether such behaviour is important in practice, I first extend the theoretical ratchet effect literature by developing a model of finite-horizon dynamic gaming. Given an environment where school-level targets depend on student prior scores, I show that the dynamic effect depends crucially on variation in the horizon, with teachers distorting their effort less when their decision affects fewer future scores within the same school. I then exploit variation in the grade span of schools to identify this effect, making use of rich educational data from North Carolina that tracks students, teachers and schools over time. I find compelling evidence of dynamic distortions using a difference-in-differences approach. The disparity in grade five teacher effort between K-5 and K-8 schools is estimated to account for between 15% and 22% of a standard deviation in the grade five score, with a similar effect obtained when comparing K-5 and K-6 schools. I then directly estimate the structural parameters of the corresponding econometric model, allowing for complementarities in production between teacher effort and student ability. Using these estimates, I carry out two counterfactual policy experiments. First, simulating a setting without any accountability scheme, the grade five score in K-5 schools would be approximately 1.25 standard deviations lower, revealing the substantial positive effects of the reform. The second experiment eliminates ratchet effects entirely, taking advantage of a key prediction of the model. Doing so results in an average grade five score that is 4.6% of a standard deviation higher, but is also around 36% more costly to implement, given that the theoretical prescription involves lowering the target, making it easier to satisfy.

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 prior student 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 useful 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

[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), Kanemoto and Macleod (1992), and Gibbons (1996) address ratchet effects under various mechanisms with limited or no commitment.

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 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 school target 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 prior student 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 and adopt is one where principals centrally coordinate and monitor teachers to maximize their school's payoff.[3] Thus the agents 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 a school. 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, so teachers will increase their effort. And in the limiting case, if the horizon consisted of 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 school'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 behaviour in a

[3]Although not explicitly considered in this paper, principals may engage in the within-school reassignment of teachers to classes according to teaching ability, in addition to influencing teacher effort.

[4]Strictly speaking, 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.

particular grade across schools with different grade spans, the model implies that schools serving fewer future grades should exert greater effort than those serving a greater number of future grades. For example, grade five teachers at K-5 schools are predicted to exert a higher level of effort than their K-8 or K-6 counterparts, leading to a positive score differential in favour 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 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.

Estimating the full difference-in-differences specification with controls, the analysis reveals substantial distortions between K-5 and K-8 schools — between 15% and 22% of a standard deviation in the grade five score in favour of the shorter grade span. These findings are consistent with the predictions of the model. The analogous distortion in grade five for the

[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.

comparison between K-5 and K-6 schools is between 21% and 30% of a standard deviation. These results are obtained without having to make overly restrictive identifying assumptions and are robust to a number of potential threats to validity, all of which involve differentially trending unobservables across school configurations unrelated to incentives. One threat arises from supply-side changes in the distribution of school configurations, potentially biasing estimates if the grade span transitions have a significant effect on the average quality of students or teachers across school types. Limiting the analysis to only those schools that do not switch configurations during the period of interest, I find that such changes are not responsible for my results. Other potential sources of bias include differential changes in peer effects and the confounding effects of parallel reforms, such as the introduction of charter schools. Given additional supporting evidence, including triple differences that compare difference-in-differences estimates across grades, I find little evidence to suggest that these threats undermine the main identification strategy.

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, 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 behaviour.

Rather than following that course, 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 prior student 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 approximately 1.25 standard deviations 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 4.6% of a standard deviation higher, but is also around 36% 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 understates the cumulative effect of ratcheting behaviour by 9.2%, thereby providing an estimate that sheds light on the usefulness of the linear approximation.

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 illustrating 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 two counterfactual policy experiments, and Section 10 concludes.

2 Prior Literature

The current research contributes to three main strands of literature. 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 behaviour 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 behaviour 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 behaviour, 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 behaviour 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, 2007) 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

[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.

provide evidence indicating that greater accountability has a positive effect on student scores.

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 (2007) reveal evidence consistent with Chicago teachers 'teaching to the distribution' of students. In addition, Jacob and Levitt (2006) 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.

Barlevy and Neal (2010) 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 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

[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.

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 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 improvement by proposing an alternative scheme to eliminate ratcheting behaviour.

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. In addition, 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

[9]Given that targets under the North Carolina accountability reform depend on prior student 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.

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 behaviour. 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 (described more fully in the next section). In that setting, incentive targets depend on prior student scores, and a single agent (the school principal) coordinates actions across grades, generating differential behaviour according to the school's grade horizon.[10] The model is related to Weitzman (1980), who 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).

[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 behaviour 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 below, for example).

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 prior student scores in North Carolina and also the fact 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 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. Thus, 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

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

[14] This modelling choice is made to focus on the core idea of ratcheting behaviour. It assumes that the principal is capable of perfectly co-ordinating her teachers. If this is not the case, then ratchet effects will 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, 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, $g = 0$ does not exist, as it represents grade five at a different school. Thus, in this case, $g \in \{1, 2, 3\}$.

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.

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 accumulated stock of skills and knowledge of a student at 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 contribution 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.[18] In addition, teacher effort is modelled exclusively as the representative teacher's contribution to the average score of her students, meaning that I abstract away from multiple tasks, such as devoting effort to disciplining students. I also initially consider

[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 modelled 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]This is a reasonable assumption to make given that the average outcome for each grade is what matters for satisfying the school-level target.

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.[19] In general, let the student's score in school s, grade configuration c, grade g and time t be given by

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

which potentially allows for teacher effort and the capital stock of the average student to interact in the production of learning. 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.[20] 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

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

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).

[19] 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.

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

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). Using the convex 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

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

which cannot be used to solve for each effort level analytically. However, the 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 rise in 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. 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 simplification 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

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

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.[21]

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 outpaces 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 $g \in \mathcal{G}_c$. 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.

[21] 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.

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.[22] Therefore, 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, since dynamic distortions should be smaller for the former type of school. 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. 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 $g,\ \forall\ g \in \mathcal{G}_c$. ■

[22] 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.

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 favour of the former type for each shared grade, and this difference should be greatest for grade five. Therefore, this grade five result is the main prediction to be tested empirically.

Proposition 3 *As an analogue 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 used 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 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.

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. This dimension 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 γ_g in place of γ and α_g in place of α. 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 modelling 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 analogue 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 given their complexity.

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 score for 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

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

where the distortion is effectively $\delta[\gamma - \alpha + 2B\theta(1 + \theta y_{scG_c-2t-1})]$.[23] 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 very involved and are omitted here.

[23] 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 $\alpha = \gamma + 2B\theta(1 + \theta\bar{y}_{cG_c-2t-1})$, 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 this section.

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 behaviour is predicted to have observationally equivalent effects on student scores as in the current formulation focusing on teacher effort. However, it is interesting to further examine the internal incentives that lead to the posited outcomes: this point is developed further in the conclusion.

4 The North Carolina Accountability Reform

Due to the fact that my identification strategy is based on the accountability scheme that North Carolina adopted in 1996, this section describes the scheme in greater detail. In particular, it establishes the following features of the scheme, captured in a stylized way in the preceding theory: school-level targets depend on the prior scores of students, the average of grade-specific targets must be satisfied to receive the bonus, and the grade span determines the number of such targets in the average.

Under the reform, students in grades three through eight are required to write standardized tests in reading and mathematics in each 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.[24] Thus, the expected growth targets are:

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

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

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.[25]

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 explain this component, 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.[26]

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

[24] 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.

[25] 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 the exposition.

[26] 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 significant that policymakers made an effort to address the problem of period-by-period noise.

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.[27] 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.[28]

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, which are the 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,[29] 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 $\hat{\beta}_2$.[30] Once estimated, these coefficients are used in all subsequent years when calculating expected gain targets. As such, they are treated as fixed. 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

[27] 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.

[28] 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.

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

[30] I have verified that this recipe produces the coefficients used by the North Carolina accountability scheme by 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.

the result of student ability; the differential ability of the prior teacher and/or school may also contribute. For instance, 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. In addition, since incentives are often in place in the prior period as well, effort is expected to vary. 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, which is an undesirable aspect of the reform, since teachers are held accountable for an outcome that they do not fully control.

Despite these shortcomings, the North Carolina accountability reform features a high-powered school-level reward scheme that conditions targets on prior scores. Moreover, the program is still in effect, meaning that dynamic distortions have had a chance to manifest themselves. If using prior scores to correct for heterogeneity actually results in dynamic gaming,[31] then this is a suitable environment for detecting such behaviour, since it should present differently depending on the grade configuration of the school considered.

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.[32] Given that the accountability reform took effect in 1997, I refer to 1994, 1995 and 1996 as pre-reform years, and 1997

[31] Survey evidence lends credence to this idea. Referring to a pilot version of the North Carolina accountability program, Heneman (1998) reports that very few Charlotte-Mecklenburg 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.

[32] I actually possess student-level data 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. For a graphical representation of the available data, see Appendix A.1.

through 2005 as post-reform years.[33] The main feature of the data set is that it contains yearly standardized test scores for each student in mathematics and reading from grades two to eight.[34] These scores are comparable across time and grades through the use of a developmental scale.[35] Using this scale and unique 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 include 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 of interest are the number of years of teaching experience and the score on the test used to obtain a teaching license. I also possess information on the type of location for each school, using seven 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 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, aggregated at the school level. 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 45% of the sample, while those who have not obtained a high school degree make up 11% of observations. Parents possessing a diploma from a trade school or community college account for a further 21% of the sample, and 24% of parents have been granted a 4-year college or graduate degree. Nearly two-thirds of North Carolina students are white, while

[33]The reform was implemented as a pilot program in 1996 for ten school districts consisting of 63 schools or approximately 4% 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. Alternatively defining 1994 and 1995 as the pre-reform period does not affect my results.

[34]'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.

[35]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 exactly in learning to 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.

slightly less than 30% are black, which is significantly higher than the national average and also higher than the state average for North Carolina.[36] In the data, the average teacher has about 13 years of teaching experience, the school-level average percentage of students qualifying for a free or reduced-price lunch is 44%, 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 approximately 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 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 that are open in the pre-reform period close or switch to a K-5 configuration early in the post-reform period. If these transitions are ignored, the remaining sample consists of relatively few K-6 schools. However, even under the most restrictive subsample, there are still 264 school-year observations for K-6 schools.

The data reveal that K-8 and K-6 schools are also disproportionately located in rural areas, where there are an average of 396 K-5 schools, 87 K-8 schools and 71 K-6 schools.[37] When compared to the average across all locales, the students at schools in such areas are more economically disadvantaged, as measured by greater participation in the free or reduced-price lunch program, have parents with lower educational attainment, and are less likely to be black.[38] Given the over-representation of the two comparison configurations (K-8 and K-6) in rural areas, where characteristics are tangibly different, controlling for the school's locale in the analysis is likely to be important.

5.2 The Impact of the Reform

Before utilizing econometric techniques to detect evidence of dynamic gaming, it is instructive to observe which patterns emerge in the raw data. There are two features that are

[36] 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).

[37] 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.

[38] Although, at 23%, the proportion of black students in rural North Carolina schools is still higher than the national average.

particularly interesting. The first one 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 disproportionately affected certain school configurations. Both of these questions can be addressed in a clear way using distributional plots.

Figures 1(a) and 1(b) are 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. 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 of the plots 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, the model would predict that the effort in this untested grade should be as low as possible to engineer a depressed 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 2 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 rougher than 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. Thus, 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. However, 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 then explore the robustness of the results to various identification threats.

6.1 Econometric Strategy

There are a number of reasons why interpreting a disparity in scores between two schools with dissimilar horizons as evidence of a ratchet effect may be ill-advised. First, the distribution of student ability may differ across schools. So if the average student in each school is not the same, then the respective school configurations will be associated with a different initial level of educational capital in the production process, leading to disparities in subsequent scores regardless of whether differential incentives exist. Similarly, if the quality of teachers, surrounding neighbourhood characteristics, or educational resources differ across school types, differential scores may be incorrectly interpreted as evidence of dynamic optimization.

Due to a variety of historical factors, it is certainly possible that such differences exist between K-5, K-6 and K-8 schools. At the beginning of the 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 enrolment pressures arising from immigration flows, K-6 and junior high schools became more prevalent as the century progressed.[39] 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, the popularity of transitionary middle schools began to wane in the 1980s and 1990s as the large and impersonal institutions were perceived to be inadequately serving their students. Later research also suggested that a higher number of school transitions was deleterious to student development.[40] In the current context, if

[39] See Juvonen *et al.* (2004) for a thorough history of the middle school in the United States.

[40] Juvonen *et al.* (2004) present survey evidence suggesting a negative psychological impact of moving,

schools are non-randomly selected to change configurations over time, as would be the case if poor or low-performing schools switched first, then K-5, K-6 and K-8 structures might not directly be comparable.

To isolate the variation in scores arising from dynamic incentives, I propose a difference-in-differences approach for my identification strategy, 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

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

Such an approach adjusts for both pre-existing disparities and shared changes 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. If families sort across neighbourhoods or teachers sort across schools, then the composition of educational production inputs might evolve over time. Table 2 presents a difference-in-differences analysis of student characteristics. The estimates suggest that, relative to K-8 schools, the proportion of students with educated parents declines and the proportion of black students rises at K-5 schools from the pre- to post-reform period. This is consistent with high-socioeconomic-status families sorting away from K-5 and into K-8 schools. Thus, failing to account for these student characteristics in the main regression would lead to downward-biased estimates. Therefore, my initial strategy combines difference-in-differences estimation with observable student, teacher and school controls X_{sgt}, which account for the measured effect of differential trends.

As there are many difference-in-differences estimates to consider, I first estimate the equation

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

Alspaugh (2001) uses cross-sectional test score data for rural Missouri students and finds that short-run achievement is lower for K-5/6-8 students than for their K-8 counterparts, and Hanushek *et al.* (2004) show there is a small negative impact on performance for students who switch schools for reasons unrelated to the grade structure of their school. More recently, Rockoff and Lockwood (2010) find large and significant negative effects of the elementary-to-middle-school transition on academic achievement using panel data in New York City.

where each ϕ is an interacted indicator variable that adjusts the score for every combination of grade, school type, and period.[41] 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

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

A finding of $\Phi_{\text{K5}-\text{K8},post-pre,g} > 0$ is interpreted as satisfying the criteria for dynamic gaming behaviour as in Proposition 1. The prediction of Proposition 2, that the magnitude of dynamic distortions is increasing in the grade, can also be tested by comparing $\Phi_{\text{K5}-\text{K8},post-pre,g}$ to $\Phi_{\text{K5}-\text{K8},post-pre,g+1}$. I now estimate these difference-in-differences objects to determine whether the data is consistent with ratcheting behaviour. After presenting the results, I thoroughly consider the potential for unmeasured trends of various kinds to threaten identification.

6.2 Results

Figures 1(a), 1(b) and 2 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 a variety of specifications, dictated by the components of the control vector X_{sgt}. These specifications are given in Table 3, 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 parental education of students and their ethnicity, and controls for the locale of the school. Specification (3) adds the proportion of those eligible for a free or reduced-price lunch, specification (4) additionally includes student exceptionality measures and the licensure test score of teachers, and specification (5) appends a control for the number of classes offered in a school per grade.

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 (5)

[41] Allowing control coefficients to vary by grade (β_g), which is a prerequisite for structurally estimating the model with transitory effort, or including school-level fixed effects does not appreciably alter the difference-in-differences results.

in particular, relative to a class of students with no parent having finished high school, a class of children with parents whose highest educational attainment is a high school diploma is predicted to have a score that is approximately 5.3 developmental scale points higher, while all students in a class having a parent with a four-year college degree extends the gain by a further 9.2 points. With respect to ethnicity, a class of black students is predicted to have a score that is nearly 10 points lower than a class of non-black students. These are large differences, as the standard deviation of the grade five score reported in Table 1 is 7.9 developmental points. The score is also positively linked to students attending a school with a lower free or reduced-price lunch participation rate, those with teachers who scored higher on their licensing test, and those attending schools with fewer numbers of classes per grade. In the case of free or reduced-price lunch participation, the difference in score between a fully participating class and one in which no student qualifies is about 5 developmental points in favour of the latter class.

For specifications (1) through (5), as defined in Table 3, and grades three through five, I transform the relevant fixed effects from equation (8) into first-difference and difference-in-differences estimates, as in equation (9). The results for K-5 and K-8 schools, and K-5 and K-6 schools are reported in Table 4 and 5, respectively. 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 (5), the pre-to-post gain in grade five scores for K-5, K-8 and K-6 schools is 10.2, 8.5 and 7.8 developmental scale points, respectively. The gains are also decreasing in the grade so that the grade three counterparts are 7.5, 7.0 and 5.6 points, respectively.[42]

The more interesting results with regard to ratchet effects are the difference-in-differences estimates. For the comparison between K-5 and K-8 schools, the difference-in-differences estimates reported in Table 4 are statistically indistinguishable from zero for each grade when no observable controls are included. However, after introducing controls, the grade four and five estimates are positive and significant, which is consistent with Proposition 1. That is, controlling for trending observables and the pre-reform outcome, the school with the shorter grade horizon (K-5) has a higher score. Moreover, although somewhat imprecise, the point estimates are also increasing in the grade, which is in keeping with the prediction of Proposition 2.

[42]From Table 1, the standard deviation of the score in both grade three and four is 8.4 points, which is actually higher than the value for grade five (7.9 points). Thus, adjusting for variation in scores, the grade three and four gains are even smaller relative to those in grade five.

The magnitude of dynamic distortions suggested by the difference-in-differences estimates is substantial. Comparing K-5 and K-8 schools, the differential effect of the scheme is estimated to be between 1.17 and 1.73 developmental scale points for grade five, depending on the control-based specification used. This is equivalent to an effect that is between 14.8% and 21.9% of a standard deviation in the grade five score. It is informative to place these results in context. In the data, a one standard deviation increase in the proportion of students with parents whose highest educational attainment is a four-year college degree, and an equivalent reduction for those with parents whose highest educational attainment is a high school diploma, raises the average score by 2.10 points. In addition, a one standard deviation decrease in the proportion of students receiving a free or reduced-price lunch is predicted to increase the test score by 1.06 points. Thus, the dynamic distortion between K-5 and K-8 schools is slightly weaker than the parental education effect and slightly stronger than the effect of reducing the proportion of students on subsidized lunches.

As with the comparison between K-5 and K-8 schools, Table 5 shows that the difference-in-differences estimates for K-5 and K-6 schools are statistically indistinguishable from zero for grade five and slightly positive and significant for grades three and four when no observable controls are included. When including controls for differentially trending observables, the difference-in-differences estimates become positive and significant for all grades, with the grade five distortion accounting for between 1.63 and 2.40 developmental scale points.[43] Thus, the main prediction of a positive disparity in grade five scores is borne out by comparing K-5 schools with both K-8 and K-6 schools.

6.3 Threats to Identification

The main challenge to the proposed identification strategy is that unobserved factors which are unrelated to incentives may vary over time. Given that this seems to be the case for observed characteristics (see Table 2), this concern cannot be easily dismissed. To address this issue, I first develop a formal condition relating observed and unobserved factors that must be satisfied for the incentive effect to be identified. I then provide specific examples of the threats I have in mind and assess the extent to which they will lead to upward bias in estimates.

[43] Although these magnitudes seem larger than for the comparison between K-5 and K-8 schools, this cannot be statistically established. This is due to the fact that there are far fewer post-reform observations for K-6 schools than for K-8 schools.

A Formal Condition

Consider two school types c and c'. Using equation (1), the grade g steady-state score for a representative school of type c is $y_{cg} = \gamma y_{cg-1} + e_{cg} + a_{cg} + u_{cg}$. Suppose that the average student in a type c school begins grade $g = 1$, the first grade with high-powered incentives attached, with an initial level of educational capital k_{c0}. Then, the pre- or post-reform score can be re-expressed as

$$y_{cg\tau} = \gamma^{g+1} k_{c0\tau} + \sum_{i=0}^{g} \gamma^{g-i}(e_{ci\tau} + a_{ci\tau} + u_{ci\tau})$$

where $\tau \in \{pre, post\}$. Defining the first difference in scores between the pre- and post-reform period as $\Delta y_{cg} \equiv y_{cg,post} - y_{cg,pre}$ and the difference-in-differences in scores between configurations as $\Delta\Delta y_{cc'g} \equiv \Delta y_{cg} - \Delta y_{c'g}$,[44] the latter quantity is

$$\Delta\Delta y_{cc'g} = \gamma^{g+1}\Delta\Delta k_{cc'0} + \sum_{i=0}^{g} \gamma^{g-i}(\Delta\Delta e_{cc'i} + \Delta\Delta a_{cc'i} + \Delta\Delta u_{cc'i})\,.$$

If the difference between types of all other inputs were time invariant ($\Delta k_{c0} = \Delta k_{c'0}$, $\Delta a_{cg} = \Delta a_{c'g}$ and $\Delta u_{cg} = \Delta u_{c'g}$, $\forall\ g$), then it would be an unbiased measure of the dynamic distortion in incentives arising from the scheme ($\Delta\Delta y_{cc'g} = \sum_{i=0}^{g} \gamma^{g-i}\Delta\Delta e_{cc'i}$).

To consider the bias associated with non-incentive inputs varying over time, define $k_{c0\tau} \equiv W'_{c\tau}\lambda_k + \xi^k_{c\tau}$ and $a_{cg\tau} = a_{c\tau} \equiv Z'_{c\tau}\lambda_a + \xi^a_{c\tau}$, where teacher ability is assumed to be identical across grades for simplicity, $W_{c\tau}$ and $Z_{c\tau}$ are observed predictors of educational capital and teacher ability, respectively, and $\xi^k_{c\tau}$ and $\xi^a_{c\tau}$ are the associated unobserved determinants. Using these decompositions, the difference-in-differences score is

$$\Delta\Delta y_{cc'g} = \sum_{i=0}^{g} \gamma^{g-i}\Delta\Delta e_{cc'i} + \Delta\Delta X'_{cc'}\beta + \Delta\Delta\nu_{cc'g}\,, \tag{10}$$

where $X_{c\tau} \equiv [W_{c\tau}\ \ Z_{c\tau}]'$, $\beta \equiv [\gamma^{g+1}\lambda_k\ \ \sum_{i=0}^{g}\gamma^{g-i}\lambda_a]'$, and $\nu_{c\tau} \equiv \gamma^{g+1}\xi^k_{c\tau} + \sum_{i=0}^{g}\gamma^{g-i}(\xi^a_{c\tau} + u_{ci\tau})$. Thus, the direction of bias associated with the simple difference-in-differences estimate with no controls depends on the sign of $\Delta\Delta X'_{cc'}\beta + \Delta\Delta\nu_{cc'g}$. If it is positive, this would lead to upward bias in the estimated distortion, while a negative value would result in downward bias. Associating c and c' with K-5 and K-8 schools, respectively, Table 2 shows that for each component of X with a positive effect on scores ($\lambda_k > 0$ or $\lambda_a > 0$), the

[44] The first difference and difference-in-differences for other quantities, such as student ability, teacher ability, and teacher effort, are defined analogously.

corresponding difference-in-differences quantity is weakly negative, while it is weakly positive for all characteristics with a negative effect. Therefore, the data suggests that $\Delta\Delta X'_{cc'}\beta < 0$. This is borne out by the results in Tables 4 and 5, where the difference-in-differences estimates strengthen as controls are added to the regression. In other words, the omission of observed characteristics seems to result in a downward-biased estimate.

If the true differential effect of incentives is positive ($\sum_{i=1}^{g} \gamma^{g-i}\Delta\Delta e_{cc'i} > 0$), as posited by the theory for $\mathcal{G}_c \subset \mathcal{G}_{c'}$, then further downward bias from the omission of unobservables would imply that the effect is stronger than estimated. The real threat to identification is if unobserved factors with a positive(negative) effect on scores shift disproportionately toward K-5(K-8) schools from the pre- to post-reform period ($\Delta\Delta\nu_{cc'g} > 0$), leading to upward bias. If this is the case, then the condition that must be satisfied for identification is

$$\Delta\Delta\nu_{cc'g} < \Delta\Delta y_{cc'g} - \Delta\Delta X'_{cc'}\beta\,.$$

Therefore, the omission of unobserved factors must either result in downward bias, in the same direction as observable characteristics, or in sufficiently small upward bias such that the estimated effect is not entirely driven by non-incentive based variation.

Specific Threats

It is impossible to know with certainty whether the preceding condition is satisfied. However, considering the most likely institutional challenges to identification can be informative. In my view, the greatest threat is associated with supply-side changes in the distribution of school configurations. During the post-reform period, North Carolina policymakers increasingly shifted toward the K-5/6-8 model.[45] If the schools were systematically selected for transition on the basis of unobserved determinants of performance, bias would result. For instance, underperformers might be chosen first due to less institutional resistance. If such schools tend to be located in disadvantaged neighbourhoods, 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 and preserving identification. While it seems likely that poorer schools would be the leading candidates for reform, it could also be the case that high-performing schools would prefer to undertake the transition if the associated benefits offset the inherent costs. If they did so, then estimates would be biased upward, which is potentially a problem for identification. However, evidence from pre-reform grade 5 scores for K-8 schools that

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

transformed and those that did not is in line with the downward bias story.[46] It seems that either it is less costly for struggling schools to make the switch or the transition is disproportionately imposed on them.

To more fully address the possibility that differential trending of supply-side changes in school configurations threatens identification, I consider two pieces of evidence. One problem is that the evolution of configurations may differ according to whether a school is located in an urban or rural area. This is likely, given that K-8 and K-6 schools are overrepresented in rural areas. This possibility is addressed by specifications (2) through (5) of Table 3 with the inclusion of school locale controls. The more general supply-side problem is addressed through a robustness check that restricts the analysis to the subset of schools that do not transition to new grade structures during the period of interest. In Table 6, I compare difference-in-differences results for the full sample of schools to subsamples that omit switches in configuration, and I additionally limit observations to schools observed in all pre- and post-reform years of the study. Interestingly, comparing K-5 and K-8 schools in grade five, the effects grow when restricting the sample and even become significant for the specification without controls. The results are also stronger for the comparison between K-5 and K-6 schools for grades three through five. Moreover, although the estimates are not statistically different between grades, the point estimates are increasing in the grade for the subsamples, which conforms to the prediction of the theory. For either comparison, the increase in estimates suggests that the bias associated with supply-side changes is downward.

Nevertheless, selection bias may remain even after restricting the analysis, due to the competitive effects of switching schools on non-switching ones, assuming schools compete with each other locally. To see why, consider the example of a district with two K-8 schools, one of which is underperforming, and the other, overperforming. 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 estimates would be biased upward. Conversely, downward bias would result if the newly converted K-5 school were perceived as being less desirable. Analyzing student migration between switching and non-switching schools might help shed light on the direction of such bias, but there is a more robust way to deal with this issue.

[46]The average pre-reform grade 5 score for K-8 schools that did and did not switch is 309.0 and 310.2 developmental scale points, respectively. This difference of 2 points is equivalent to 32% of a standard deviation in the grade 5 score for all K-8 schools.

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 of the difference between such estimates for any two grades. Table 7 presents results for K-5 and K-8 schools, and K-5 and K-6 schools. The main estimates of interest compare grade five to four. For K-5 and K-8 schools, the estimates are positive and significant for the full sample and each subsample of non-switchers. The triple difference is also positive and significant for the full sample of K-5 and K-6 schools. 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 competition does not affect scores differentially by grade, then the remaining supply-side selection bias is likely to be addressed by this robustness check.

Having attended to supply-side issues, the disparity in unobservable student or teacher ability between configurations may also evolve over time due to demand-side sorting by households or teachers. However, the stories underlying such an effect are not obvious. One possibility is that economic opportunities in rural areas disproportionately decrease for low ability households, resulting in increased migration of those families to urban centres. If one type of school is overrepresented in cities, then estimates would be biased. Bias arising from this specific story can be dealt with by controlling for school locale, as in the preliminary supply-side analysis. The results are robust to such controls. Beyond that, student and teacher transfers between school types can be analyzed to see if sorting alters the distribution of ability based on observable characteristics across configurations and over time. If such an analysis suggests a downward bias and unobservables operate in a similar way, then dynamic distortions will be identified. However, there is a much simpler way to deal with demand-side bias. Given that household or teacher sorting is unlikely to differ by grade, the triple-differences analysis addresses the threat.

An additional, but more secondary, threat to identification is that differences in production may vary over time through, for instance, evolving peer effects. Older students in K-8 schools are likely to have an effect on younger students that has no analogue in K-5 schools.[47] If this generally deleterious effect changes from the pre- to post-reform period, then estimates of the ratchet effect will be biased. It is possible that older students become less of a negative influence on their younger peers over time. Jacobson (2004) documents a national trend

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

of declining soft drug use among teenagers between 1997 and 2000 that supports this idea. If this is informative about 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.

On the other hand, upward bias would result if the trend were reversed. 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.[48] Therefore, if the positive difference-in-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 8 presents difference-in-differences 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 or equal to half of the combined effect for all estimates that are significant.

The 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.[49] If charter schools cause the average scores of neighbouring 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. Yet 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 grade five students. Beginning in 2001, fifth graders were required to satisfy a certain threshold of performance in order to be promoted to the sixth grade.[50] For this reform to bias my results, it would need to affect students differentially by school configuration. While it is not clear why this would occur, 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, while upward bias would result if the opposite were true, where a stronger response might arise if a greater percentage of grade five students were marginal. Fortunately, the identification strategy can handle either type of bias, since the

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

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

[50] See Cooley (2010) for a more in-depth explanation of the reform.

student accountability reform only affects fifth-grade students and difference-in-differences comparisons are performed for grades three through five. Moreover, given that the introductions of each reform are not coincident, it is possible to isolate the distortion for grade five by conducting the analysis for subsets of the post-reform period. The results are robust to such restrictions.

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. Given that the model strongly tracks the data, 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.[51] Using $\Phi_{cc'g} = \Delta\Delta y_{cc'g} - \Delta\Delta X'_{cc'}\beta$ and assuming the difference-in-differences of the score, adjusted for observable characteristics, is an unbiased measure of the distortion ($\Delta\Delta\nu_{cc'g} = 0$), equation (10) 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, respectively, so that $G_c = G = 3$. From the model, the first-order conditions for the simplifying

[51] 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 it is potentially interesting, it is unlikely to be of first-order importance, given the aggregated level at which the analysis occurs.

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

$$\delta\gamma\Delta\Delta e_{K5,K6,G} = \Delta\Delta e_{K5,K6,G-1} = \delta^2\gamma B(\alpha-\gamma)\,.^{52} \tag{14}$$

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 2 equations with the 2 unknowns γ and B, assuming α and δ are given.[53] 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 about 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 corresponding difference-in-differences expressions are

[52] 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.

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

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

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 Structural 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 $\alpha = 0.924$,[54] and assume an inter-temporal depreciation parameter δ of 0.9.[55] Table 9 presents the structural estimates for each configuration comparison using specification (5) with full controls. For the comparison between K-5 and K-6 schools, the growth parameter γ is estimated to be 0.56, the myopic parameter B is estimated to be 4.30, 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.[56] Therefore,

[54] 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 $\alpha = 1.05 \times 0.88 = 0.924$.

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

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

the estimate 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, this 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 by 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 10 presents the structural estimates of the transitory model, using full controls. 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 directly determined 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 these for grade four and five is 0.805, 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 estimates for ω and B are both insignificant, but the point estimates are each worth discussing. The estimate for B is 2.73, which is of the same magnitude as found in the fully persistent analysis. The estimate for ω from the structural analysis is 0.50. This is an interesting result, since it is in keeping with the previously mentioned upper bound of teacher effects found in the literature.

8 Structural Estimation with Nonlinear Technology

Although reasonably state-of-the-art 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 prior student score, which is conveniently represented by the production function in equation (5). Doing so allows me to structurally test this model against the simpler linear one, by ascertaining whether the latter can be statistically rejected.[57] 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. 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 between configurations and grades. The estimation problem is to select the parameter values that maximize the log-likelihood function

$$\mathcal{L}(\gamma, \theta, B, \sigma^2) = \sum_{t=1}^{T} \sum_{s=1}^{S} \sum_{g \in G_c} ln(\varphi(u_{scgt}; \gamma, \theta, B, \sigma^2)\,,^{58} \tag{16}$$

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).[59]

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 fixed effects at the configuration-grade level, which is all that is really required to identify effort at the 'horizon' level. This is done over all available time periods, since period-specific effects cannot be separately identified from period-by-period effort levels. This results in an analysis which controls for differences in the level of ability between configurations and grades, but does not account for common trends over time as in the

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

[58] 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.

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

difference-in-differences linear approach. The greater flexibility associated with this is what allows identification of the additional interaction parameter θ. After controlling for fixed effects, there are essentially three types of school in the likelihood function: schools serving the final grade G, the second-from-last 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 behaviour 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 between configurations and grades. I utilize the quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) gradient method to solve the unconstrained 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 350 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 11 for $\alpha = 0.924$ and $\delta = 0.9$.[60] Confidence intervals at the 90 and 95 percent level are reported. The estimate for γ is 0.868, while B is estimated to be 1.17. 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 0.0024, rejecting the more restrictive linear technology hypothesis at the 90 percent level and nearly doing so 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. Although Proposition 8 implies that the distortion should be attenuated for

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

$\theta > 0$, γ is also smaller under the nonlinear specification, which means that the overall distortion (which is approximately $\delta[\gamma - \alpha + 2B\theta(1 + \theta y_{scG_c-2t-1})]$) for a given prior score $y_{scg-1t-1}$ is larger under the less restrictive nonlinear specification. Thus, the distortion between grades is underestimated by assuming the technology is linear.

9 Policy Experiments

Although the structural estimates provide useful insight into the technology that underlies the learning process, one reason for going beyond a reduced-form analysis of ratchet effects is to carry out illuminating policy experiments.

The first experiment involves exploring a counterfactual world in which the reform was never enacted. This sheds light on the true 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 top panel of Table 12. Using the general nonlinear structural estimates, $\gamma = 0.868$, $B = 1.17$ and $\theta = 0.0024$, the resulting cumulative grade five score at the average K-5 school is approximately 1.25 standard deviations lower than the actual level that is observed. Thus, in keeping with the descriptive evidence, the reform had a substantial effect on student achievement.

The second experiment uses the theoretical prediction that the ratchet effect is eliminated by choosing the target $\alpha = \gamma + 2B\theta(1 + \theta\bar{y}_{cG_c-2t-1})$, where $\bar{y}_{cG_c-2t-1}$. 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 increases in early effort have a compounding effect on the grade five score due to the role of the production technology. The results are presented in the bottom panel of Table 12. The cumulative effect of eliminating ratcheting behaviour at the average K-5 school is a 4.6% of a standard deviation increase in the grade five score. However, such a scheme is about 36% more costly to implement, as the target α is lowered to thwart dynamic gaming, making it easier to satisfy.

There are alternative ways to formulate the relevant non-linearities, with the chosen specification proving to be very analytically tractable for the purposes of comparing the preceding counterfactuals with their linear counterparts. These comparisons are found in Table 12.

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 approximately one standard deviation. Counterfactually eliminating the distortion instead, I find that the cumulative increase in the grade five score is 4.2% of a standard deviation. Comparing this result to the more flexible nonlinear result, the linear simplification underestimates the total distortion by 9.2%. On the other hand, the linear scheme without distortions is estimated to cost nearly 39% more, which is overestimated compared to the nonlinear figure.

The comparison between the nonlinear and linear technologies is interesting. The more general specification yields counterfactual results that are substantially different from the restricted linear case. In particular, the cumulative effect of the reform is understated by about 20%, while the cumulative effect of ratcheting behaviour is understated by about 9%. These nontrivial disparities suggest caution may be 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 corresponding performance targets is that they adjust for unobserved heterogeneity in scholastic inputs. However, a rich dynamic incentive theory literature predicts that the inherent intertemporal 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 crucial for policymakers to understand whether ratchet effects arise in practice and if so, how much they distort outcomes. Yet 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 the literature is that existing theoretical formulations do not provide a clear prediction as to where one might look for such dynamic effects, an important ingredient for forming a plausible identification strategy. In light of this, I 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 difference-in-differences strategy, I find substantial evidence of such effects, with distortions ranging between 15% and 22% of a standard deviation in the grade five score. These dynamic results are an important addition to 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 approximately 1.25 standard deviations 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 4.6% of a standard deviation higher in the absence of ratchet effects, but would also be about 36% more expensive to implement, making it a relatively undesirable remedy for policymakers.

The results of this analysis suggest several avenues for further research. Given the distortions associated with ratcheting behaviour, I am in the process of isolating the channels through which the effect operates, the leading candidates being differential effort exertion and reassignment of teachers across classrooms. To that end, I am working on a richer model that augments the effort decision with endogenous teacher assignments by the school principal, based on teacher quality. This I intend to estimate with the same rich North Carolina data, though making use of additional information on teacher assignments. The dynamic framework that I have developed also points to a procedure for inferring idiosyncratic effort — typically a challenging task — that I plan to implement in related work combining theoretical modelling and structural estimation. Given the magnitude of the dynamic distortions I find, in a more personnel economics vein, I am keen to investigate possible management practices that may help to account for the measured effects. With this in mind, I would like to conduct a survey of school principals and district officials in North Carolina in light of the state accountability scheme, as a supplement to my econometric analysis.

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, J. 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 (2010), "Pay for Percentile," working paper, *http://sites.google.com/site/dereknealsresearch/october10.pdf*.

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 (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

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. 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 Behavioural Sciences*, 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*, forthcoming.

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	290.2	8.4	250.3	316.9
Grade 4	302.7	8.4	261.0	329.1
Grade 5	314.5	7.9	277.0	340.1
Grade 6	321.6	8.2	276.7	345.8
Grade 7	331.1	7.6	294.5	353.8
Grade 8	337.1	7.4	298.1	357.9
Student - Parental Education:				
No High School	0.11	0.10	0	1
High School Graduate	0.45	0.17	0	1
Trade School	0.09	0.09	0	1
Community College	0.12	0.09	0	0.7
4-Year College	0.19	0.15	0	1
Graduate Degree	0.05	0.07	0	1
Student - Ethnicity:				
White	0.63	0.29	0	1
Black	0.28	0.26	0	1
Other	0.09	0.14	0	1
Student - Exceptionality:				
Learning Impairment*	0.12	0.07	0	1
No Special Label*	0.76	0.12	0	1
Gifted*	0.13	0.11	0	1
Teacher:				
Experience*	13.2	6.5	0	42
License Test Score*	0.01	0.58	-3.42	2.52
School - Locale:				
Large City	0.05	0.22	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
Large Town	0.01	0.10	0	1
Small Town	0.13	0.34	0	1
Rural	0.42	0.49	0	1
School - Other:				
% Free or Reduced-Price Lunch*	0.44	0.22	0	1
Avg. No. of Classes (Gr. 3-5)*	3.5	1.4	0	12

Note: Statistics averaged at the school level from 1994 to 2005 (* indicates no data for 1994). Student and school location categories are both mutually exclusive and exhaustive.

Table 2: Evolution of Student Controls ((K-5 − K-8)$_{post-pre}$)

Dependent Variable	Grade 3	Grade 4	Grade 5
Pared - No HS	0.01	0.01	0.01
	(0.01)	(0.01)	(0.01)
Pared - HS Graduate	0.03*	0.05**	0.03†
	(0.01)	(0.01)	(0.01)
Pared - Trade School	0.00	0.00	0.00
	(0.01)	(0.01)	(0.00)
Pared - Community College	-0.02*	-0.02*	-0.02*
	(0.01)	(0.01)	(0.01)
Pared - 4-year College	-0.01	-0.03**	-0.01
	(0.01)	(0.01)	(0.01)
Pared - Graduate Degree	-0.01**	-0.01**	-0.01**
	(0.00)	(0.00)	(0.00)
Ethnic - Black	0.00	0.02	0.02†
	(0.01)	(0.01)	(0.01)

Standard errors adjusted for clustering at school level are reported in parenthesis.
Significance levels : ** : 1% * : 5% † : 10%

Table 3: Reduced-Form Specifications

Dependent Variable: Combined Mathematics and Reading Score					
Regressor\Specification:	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Student - Parental Education:					
No High School		-31.34**	-28.27**	-21.72**	-21.50**
		(1.12)	(1.18)	(0.93)	(0.91)
High School Graduate		-24.16**	-22.42**	-16.40**	-16.20**
		(0.96)	(0.93)	(0.73)	(0.72)
Trade School		-20.57**	-19.53**	-14.55**	-14.33**
		(1.08)	(1.06)	(0.84)	(0.83)
Community College		-21.07**	-20.09**	-14.31**	-14.14**
		(1.10)	(1.07)	(0.85)	(0.84)
4-Year College		-9.48**	-10.50**	-7.48**	-7.02**
		(1.09)	(1.05)	(0.80)	(0.78)
Student - Ethnic - Black		-12.13**	-9.97**	-9.87**	-9.80**
		(0.34)	(0.51)	(0.48)	(0.48)
Student - Exceptionality:					
Learning Impairment				-10.26**	-10.45**
				(0.83)	(0.82)
Gifted/Exceptional				10.43**	10.35**
				(0.55)	(0.54)
School - % Free Lunch Eligible			-5.00**	-4.76**	-5.12**
			(0.71)	(0.64)	(0.64)
Teacher - License Test Score				0.46**	0.46**
				(0.09)	(0.09)
Number of Classes by Grade					-0.14**
					(0.03)
Constant	310.5**	336.4**	337.6**	343.2**	343.4**
	(0.9)	(1.3)	(1.3)	(4.4)	(4.4)
School Locale Controls?	No	Yes	Yes	Yes	Yes
R^2	0.825	0.918	0.921	0.946	0.946
Observations	51591	51092	45439	44108	44108

Note: This table defines five 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 (year × type × grade × experience), which are used to construct the first-differences and difference-in-differences estimates reported in Tables 4 and 5. The analysis is done for the years 1994 through 2005, with the number of observations declining as regressors with missing values are added.

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

Significance levels : ** : 1% * : 5%

Table 4: Reduced-Form Results for K-5 and K-8

Specification:	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Grade 5					
$\Phi_{\text{K5},post-pre,5}$	9.11**	9.73**	10.32**	10.14**	10.23**
	(0.22)	(0.16)	(0.20)	(0.18)	(0.18)
$\Phi_{\text{K8},post-pre,5}$	8.95**	8.57**	8.71**	8.44**	8.49**
	(0.42)	(0.38)	(0.46)	(0.42)	(0.42)
$\Phi_{\text{K5-K8},post-pre,5}$	0.16	1.17**	1.60**	1.70**	1.73**
	(0.48)	(0.41)	(0.48)	(0.44)	(0.44)
Grade 4					
$\Phi_{\text{K5},post-pre,4}$	8.36**	8.74**	9.08**	8.87**	8.94**
	(0.19)	(0.13)	(0.17)	(0.16)	(0.16)
$\Phi_{\text{K8},post-pre,4}$	8.46**	7.94**	8.39**	8.22**	8.28**
	(0.41)	(0.38)	(0.40)	(0.37)	(0.37)
$\Phi_{\text{K5-K8},post-pre,4}$	-0.10	0.80*	0.69†	0.65†	0.65†
	(0.46)	(0.40)	(0.41)	(0.39)	(0.39)
Grade 3					
$\Phi_{\text{K5},post-pre,3}$	6.84**	7.04**	7.42**	7.44**	7.53**
	(0.20)	(0.15)	(0.18)	(0.17)	(0.17)
$\Phi_{\text{K8},post-pre,3}$	6.78**	6.48**	6.55**	6.90**	6.97**
	(0.44)	(0.38)	(0.39)	(0.39)	(0.39)
$\Phi_{\text{K5-K8},post-pre,3}$	0.06	0.56	0.86*	0.54	0.56
	(0.49)	(0.41)	(0.41)	(0.41)	(0.41)

Note: For each specification defined in Table 3 and according to grade, this table reports first-differences and difference-in-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 : ** : 1% * : 5% † : 10%

Table 5: Reduced-Form Results for K-5 and K-6

Specification:	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
Grade 5					
$\Phi_{\text{K5},post-pre,5}$	9.11**	9.73**	10.32**	10.14**	10.23**
	(0.22)	(0.16)	(0.20)	(0.18)	(0.18)
$\Phi_{\text{K8},post-pre,5}$	8.41**	8.10**	8.17**	7.76**	7.83**
	(0.46)	(0.32)	(0.35)	(0.34)	(0.34)
$\Phi_{\text{K5-K8},post-pre,5}$	0.70	1.63**	2.15**	2.38**	2.40**
	(0.53)	(0.36)	(0.39)	(0.37)	(0.37)
Grade 4					
$\Phi_{\text{K5},post-pre,4}$	8.36**	8.74**	9.08**	8.87**	8.94**
	(0.19)	(0.13)	(0.17)	(0.16)	(0.16)
$\Phi_{\text{K8},post-pre,4}$	7.30**	6.74**	7.27**	7.09**	7.16**
	(0.49)	(0.34)	(0.35)	(0.33)	(0.33)
$\Phi_{\text{K5-K8},post-pre,4}$	1.06*	1.99**	1.81**	1.78**	1.78**
	(0.54)	(0.36)	(0.36)	(0.35)	(0.35)
Grade 3					
$\Phi_{\text{K5},post-pre,3}$	6.84**	7.04**	7.42**	7.44**	7.53**
	(0.20)	(0.15)	(0.18)	(0.17)	(0.17)
$\Phi_{\text{K8},post-pre,3}$	5.74**	5.03**	5.22**	5.56**	5.63**
	(0.48)	(0.35)	(0.37)	(0.36)	(0.36)
$\Phi_{\text{K5-K8},post-pre,3}$	1.10*	2.01**	2.19**	1.88**	1.91**
	(0.53)	(0.37)	(0.38)	(0.38)	(0.38)

Note: For each specification defined in Table 3 and according to grade, this table reports first-differences and difference-in-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 : ** : 1% * : 5%

Table 6: Restricted-Sample Robustness Check

	K-5 vs. K-8		K-5 vs. K-6	
Specification:	**(1)**	**(5)**	**(1)**	**(5)**
Grade 5 Diff-in-Diff				
Full Sample	0.16	1.73**	0.70	2.40**
	(0.48)	(0.44)	(0.53)	(0.37)
No Switchers	1.12*	2.10**	0.09	2.92**
	(0.52)	(0.54)	(0.75)	(0.53)
No Switch - All Years	0.71	1.97**	-0.90	2.36**
	(0.51)	(0.55)	(0.78)	(0.53)
Grade 4 Diff-in-Diff				
Full Sample	-0.10	0.65†	1.06*	1.78**
	(0.46)	(0.39)	(0.54)	(0.35)
No Switchers	0.41	0.47	1.59†	2.61**
	(0.50)	(0.43)	(0.91)	(0.59)
No Switch - All Years	0.02	0.46	0.67	2.09**
	(0.48)	(0.43)	(0.80)	(0.57)
Grade 3 Diff-in-Diff				
Full Sample	0.06	0.56	1.10*	1.91**
	(0.49)	(0.41)	(0.53)	(0.38)
No Switchers	0.05	0.31	1.07	1.95*
	(0.55)	(0.48)	(0.91)	(0.88)
No Switch - All Years	-0.26	0.30	0.15	1.85†
	(0.50)	(0.50)	(1.04)	(1.04)

Note: For the specification without any and with full controls, this table reports robustness checks for the difference-in-differences estimates in each grade by comparing the full sample results to those for restricted subsamples. As before, the estimates are constructed from joint F-tests of the interaction dummies included in the relevant regression.

Standard errors adjusted for clustering at school level are reported in parenthesis.
Significance levels : ** : 1% * : 5% † : 10%

Table 7: Supporting Evidence - Triple Difference

	K-5 vs. K-8	**K-5 vs. K-6**
$\Phi_{\text{K5-K8/K6},post-pre,5-4}$		
Full Sample	1.08*	0.62†
	(0.48)	(0.36)
No Switchers	1.63**	0.32
	(0.60)	(0.54)
No Switch - All Years	1.51*	0.26
	(0.62)	(0.62)
$\Phi_{\text{K5-K8/K6},post-pre,4-3}$		
Full Sample	0.09	-0.13
	(0.51)	(0.37)
No Switchers	0.15	0.65
	(0.48)	(0.90)
No Switch - All Years	0.16	0.24
	(0.48)	(0.94)

Note: This table presents triple-difference results for the full samples and two subsamples of the data by taking the difference between difference-in-differences estimates across grades. All triple differences are determined using specification (5) with full controls and the original difference-in-differences 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: ** : 1% * : 5% † : 10%

Table 8: Supporting Evidence - Breakdown by Subject

	K-5 vs. K-8		K-5 vs. K-6	
Specification:	**(1)**	**(5)**	**(1)**	**(5)**
Grade 5 Diff-in-Diff				
Φ Combined	1.12*	2.10**	0.09	2.92**
	(0.52)	(0.54)	(0.75)	(0.53)
Φ Mathematics	0.84*	1.35**	0.01	1.46**
	(0.35)	(0.38)	(0.47)	(0.38)
Grade 4 Diff-in-Diff				
Φ Combined	0.41	0.47	1.59†	2.61**
	(0.50)	(0.43)	(0.91)	(0.59)
Φ Mathematics	0.18	0.19	0.98†	1.72**
	(0.31)	(0.27)	(0.57)	(0.41)
Grade 3 Diff-in-Diff				
Φ Combined	0.05	0.31	1.07	1.95*
	(0.55)	(0.48)	(0.91)	(0.88)
Φ Mathematics	-0.12	0.02	0.70	1.27*
	(0.32)	(0.30)	(0.55)	(0.55)

Note: This table compares the difference-in-differences estimates for the combined score to those for mathematics. The estimates are constructed from joint F-tests of the interaction dummies included in the relevant regression for the subsample of schools that do not switch configuration during the period of analysis. The coefficient for reading is simply the difference between Φ Combined and Φ Mathematics.
Standard errors adjusted for clustering at school level are reported in parenthesis.
Significance levels : ** : 1% * : 5% † : 10%

Table 9: Structural Results - Fully Persistent Linear Technology

	K-5 vs. K-8	**K-5 vs. K-6**
γ	0.34	0.56**
	(0.21)	(0.16)
B	2.05**	4.30**
	(0.53)	(0.99)

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, using $\delta = 0.9$ and $\alpha = 0.924$.
Standard errors adjusted for clustering at school level are reported in parenthesis.
Significance levels: ** : 1% * : 5%

Table 10: Structural Results - Linear Tech. with Transitory Effort (K-5 vs. K-6)

γ_6	0.66^{**}
	(0.19)
γ_5	0.79^{**}
	(0.12)
γ_4	0.72^{**}
	(0.10)
γ_3	0.84^{**}
	(0.06)
γ_2	0.77^{**}
	(0.07)
ω	0.50
	(0.56)
B	2.73
	(3.01)

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, using $\delta = 0.9$ and $\alpha = 0.924$. Standard errors adjusted for clustering at the school level are reported in parenthesis.
Sig. levels: $**$: 1% $*$: 5%

Table 11: Structural Results - Nonlinear Technology

	2.5%	**5%**	**Pt. Est.**	**95%**	**97.5%**
γ	0.8630	0.8638	0.8684	0.8740	0.8756
θ	-0.0001	0.0001	0.0024	0.0057	0.0060
B	0.3809	0.4093	1.1729	2.6625	3.0673
σ	4.2087	4.2179	4.2550	4.2958	4.3042

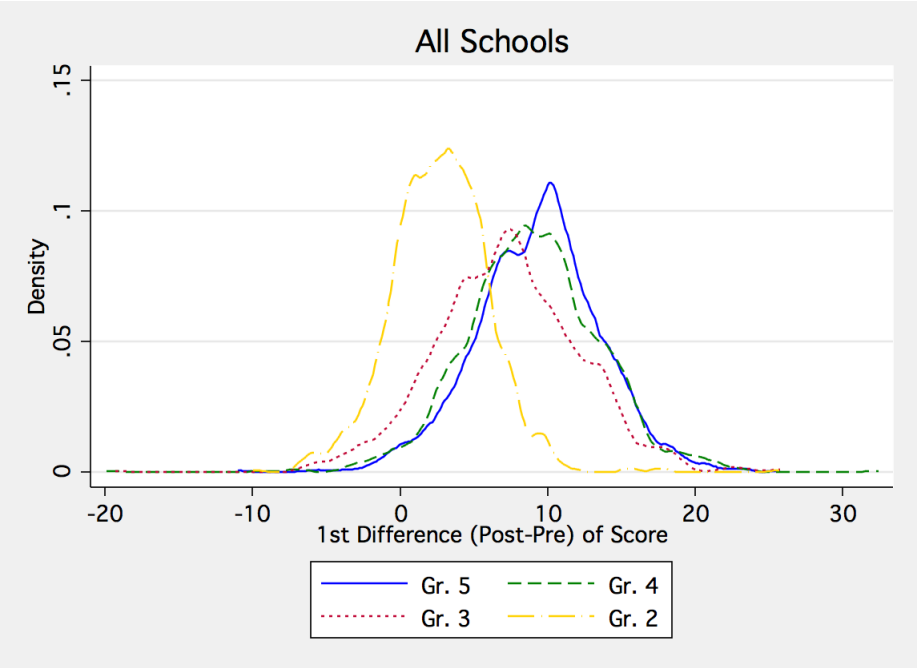
Note: This table presents structural parameter estimates for the model with nonlinear production technology. Parameters are estimated using maximum-likelihood estimation for $\delta = 0.9$ and $\alpha = 0.924$. Confidence bounds are obtained from the relevant percentiles of the bootstrap distribution, computed using 350 draws from the underlying error structure.

Table 12: Counterfactual Simulations

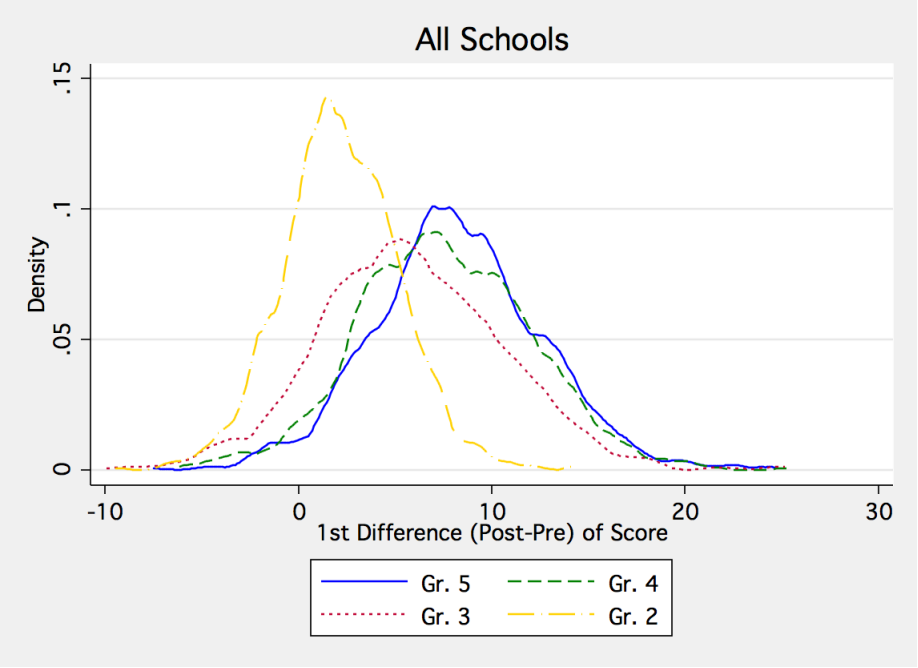
	Nonlinear	**Linear**
No Reform		
counterfactual score	307.672	309.059
actual score	316.424	316.424
Δ in score	-8.752	-7.365
Δ as % of st. dev.	-126%	-106%
No Distortion		
counterfactual score	316.742	316.713
actual score	316.424	316.424
Δ in score	0.319	0.290
Δ as % of st. dev.	4.59%	4.17%
% Δ in cost	36.17%	38.85%

Note: Analysis for K-5 schools. The actual mean and standard deviation of the average post-reform grade five score is 316.424 and 6.941, respectively. The nonlinear simulation uses estimates $\gamma = 0.868$ and $B = 1.17$ and $\theta = 0.0024$, 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 $\delta = 0.9$.

Figure 1: Density of First-Differenced Scores By Grade

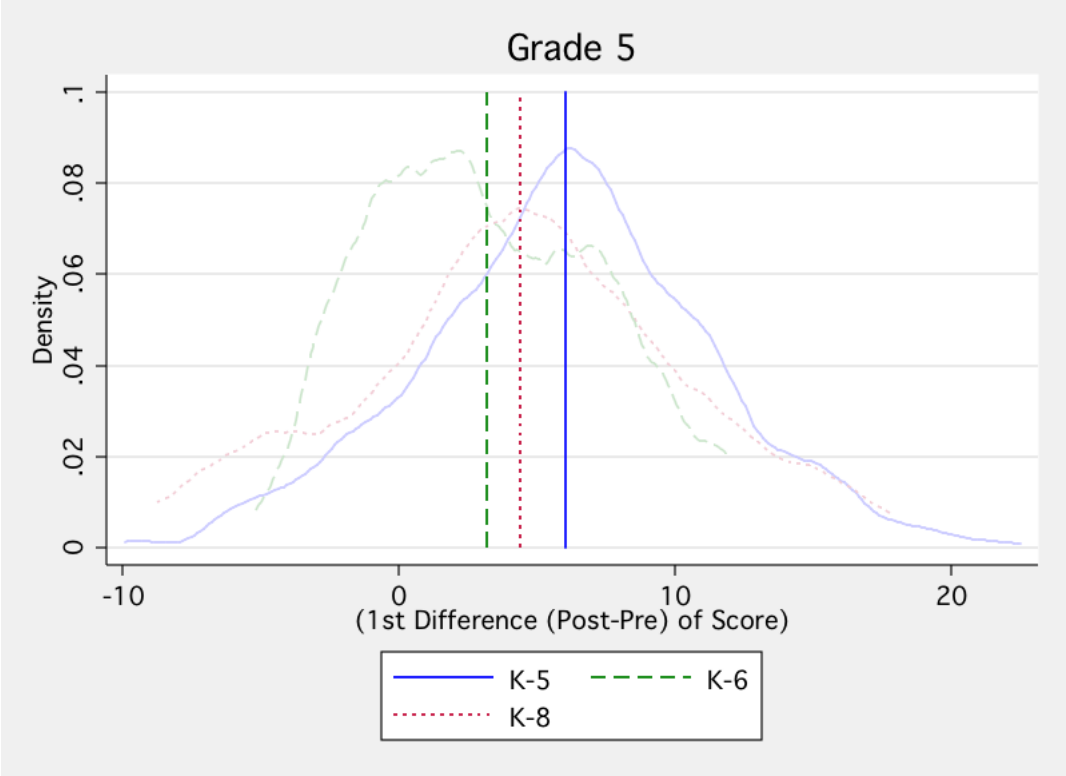


(a) Raw Score



(b) Adjusted Score

Figure 2: Grade 5 Distribution of First-Differenced Scores By Configuration



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.