Assessing the Role of Teacher - Student Interactions*

Esteban M. Aucejo†

November 10, 2011

Abstract

Teacher effectiveness is generally characterized by a single effect that is common across students. However, educators are multi-task agents that choose how to allocate their efforts among pupils. Some teachers may target their courses towards the top students in the class while others to the bottom, leading to different complementarity effects. Moreover, the introduction of accountability programs, such as No Child Left Behind (NCLB), could induce a reallocation of teacher's efforts, affecting the dynamics of student-teacher interactions. This study shows that the role of complementarities is key from a policy perspective. In this regard, an analytical framework and a novel iterative algorithm are implemented in order to characterize and quantify these effects. Results indicate that interaction effects played a crucial role in shaping the distribution of student achievement, especially after the implementation of NCLB. While more than half of the total gains in test scores experienced by the bottom third of the student achievement distribution post NCLB are due to adjustments in teacher-student complementarities, those with the very highest abilities have seen decreases in their performance.

Keywords: Interactions; Teachers; Students; Complementarities

JEL Classification Codes: I2, I20, I21.

*I am deeply grateful to Peter Arcidiacono, Pat Bayer, Federico Bugni, Joe Hotz and Arnaud Maurel for their encouragement and support. I also thank Nir Jaimovich, Hugh Macartney, Seth Sanders, Andrew Sweeting, Duncan Thomas, Christopher Timmins, Daniel Xu and the participants at Duke Seminars. All remaining errors are my own.

†Department of Economics, Duke University. Email: esteban.aucejo@duke.edu.

1 Introduction

During the last decade the federal government and many states have taken a series of steps to hold schools accountable for the performance of their students. Namely, these policies rely on student testing to measure success and failure[1]. For instance, in 2001, the No Child Left Behind (NCLB) program established that schools are subject to sanctions when the proportion of students scoring above a state-specific proficiency threshold falls below mandated levels[2]. In a similar vein, Florida has recently passed the *merit pay* bill, which eliminates tenure for new hired teachers and ties a portion of teacher pay to student performance on tests, rather than totally on principal evaluations[3].

The implementation of these policies has drawn the attention of many social scientists, leading to a prolific literature in economics on teacher effectiveness. Prominent examples include Rockoff (2004), Hanushek et al. (2005), Aaronson et al. (2007), Jacob et al. (2009), McCaffrey et al. (2009), among others. However, most articles have characterized teacher effectiveness by a single fixed effect. Therefore, the implicit assumption is that teachers have homogeneous effects across their students. In this study, teachers are allowed to have different treatment effects for different types of students.

This paper recognizes and studies the fact that educators have heterogeneous abilities and perform multi-task jobs[4]. For instance, teachers may have different preferences on how to set course level difficulty; some teachers may prefer to target course curricula based on high ability students; therefore, their capability to transmit knowledge to those pupils on the left tail of the achievement distribution may be lower. Moreover, preferences can be endogenous to classroom characteristics, implying that educators may adjust (every academic year) their teaching strategies based on, for example, the mean student in the classroom. In this regard, the aim of this article is to bring to the center of the analysis the role that heterogeneous teacher effects may have on students' outcomes, and to study the mechanisms on how these effects operate when different accountability policies are in place.

[1] However, these programs differ by the type of mechanism (i.e."carrot" or "stick") that has been used to improve low-performing schools.

[2] In addition, the schools that present high proportion of students in any of several demographic, socioeconomic, or linguistic subgroups below mandated levels are also subject to sanctions.

[3] See Kane et al. (2002) and Hanushek et al. (2004) for a description of the different programs in the US.

[4] This implies that teachers can choose how to allocate their efforts across students.

To explicitly define the notion of heterogeneous teacher effects, an analytical framework that characterizes the nature of these effects by the degree of teacher - student interactions is developed. More specifically, the analytical model considers educators as multi-task agents that choose how to allocate their efforts among different students[5]. To illustrate this point three examples are provided. Figure 1 shows the case where teachers only have homogeneous effects across students, as is often assumed. Here, effectiveness (i.e. value added) of teacher A and B is not a function of students characteristics, where A is a better teacher than B for all students. Figure 2 presents an example of different teacher effectiveness for the same group of students conditional on having three teachers with different heterogeneous effects. In this case, the three educators are equally effective for the mean student in the classroom, being the value added of teacher D the same across students. However, teacher C can be described, for example, by a lower intercept but a higher slope than teacher E; pointing out the likely situation where some teachers (e.g. E) are better suited for low performing students (e.g. student 1), while others (e.g. C) constitute a better match for high performing ones (e.g. student 2).

Moreover, the analytical framework makes it possible to test whether educators adjust their intercepts and/or slopes based on classroom average characteristics or as a response to the implementation of accountability policies[6]. For instance, the introduction of NCLB (i.e. program that intends to improve the performance of lagging students) could have affected teachers optimal decisions by incentivizing them to increase their intercepts and decrease their slopes, as is illustrated in Figure 3. Here, teacher C shifts her focus more toward low performing students at the expense of high performing students.

The econometric strategy of this paper involves the estimation of additive and interactive fixed effects (i.e. teacher intercepts and slopes). The identification and consistency properties of these types of models are not trivial when the number of observations per student is small. Basically, the incidental parameters may spoil the estimates of the parameters of interest. In this regard, Lee et al. (1993), Ahn et al. (2001), and Bai (2009) have studied the identification conditions of interactive fixed effect models. However, this manuscript distinguishes from them in several aspects. First, Lee et al. (1993) and Ahn et al. (2001) only consider the case where the interactive effect is

[5] As Holmstrom and Milgrom (1991) point out, teachers perform complex jobs involving several tasks; therefore characterizing them as multi-task agents is suitable.

[6] Macartney (2011) shows how teachers responded to the introduction of a pay per performance scheme in North Carolina.

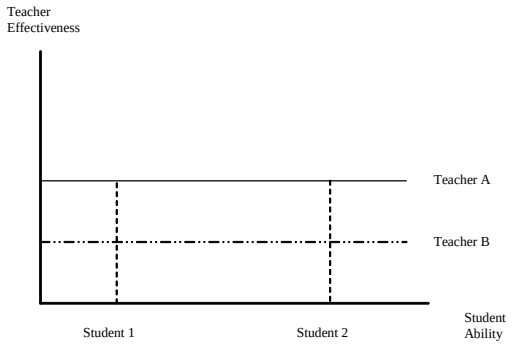


Figure 1: Characterization of teachers effectiveness (i.e. value added) when only considering homogeneous effects across students.

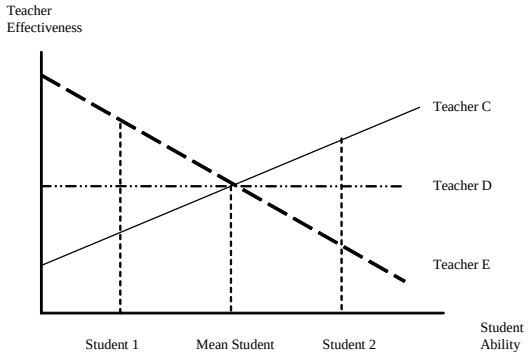


Figure 2: Characterization of teachers effectiveness (i.e. value added) based on their intercepts (i.e. homogeneous effects) and slopes (i.e. interaction/complementarity effects).

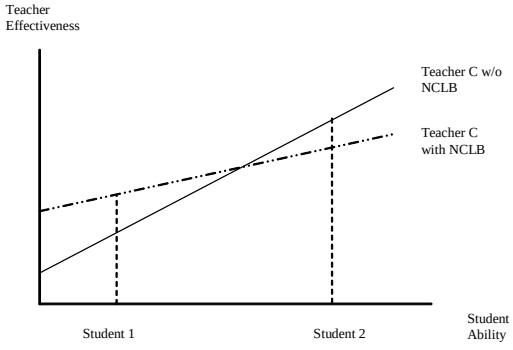


Figure 3: Example of how a given teacher could have responded to the implementation of an accountability program (e.g. NCLB).

the same for all individuals in the sample, and second their models are focused on one component models. Third, Bai (2009) works in the context of large T, while in this study $T = 3$ which is the maximum number of observations per student. Here, it is shown that after concentrating out the vector of unobserved students abilities of the original least squares problem, then it is possible to prove that consistent estimators of teachers intercepts and slopes can be recovered.

Estimation requires recovering separate intercepts and slopes for each teacher. Given the large number of parameters to be obtained and the fact that within-group transformation fails to purge interactive fixed effects, a novel iterative algorithm is applied. In the spirit of Lee et al. (1993), and Arcidiacono et al. (2011), the algorithm toggles between estimating student fixed effects and teachers slopes and intercepts. Each iteration lowers the sum of squared errors, with a fixed point reached at the nonlinear least squares solution to the full problem.

Determining the role of teacher - student interaction effects is key from a policy perspective given that it sheds light on the mechanisms behind the outcomes of different accountability policies. For example, establishing the degree on which teachers reallocate efforts in favor of some students and/or increase their overall efforts after the introduction of a new scheme of incentives can be crucial to improve policy designs. Moreover, the presence of teacher - student interactions may raise questions about the type of incentives that NCLB or pay per performance programs may trigger not only at the teacher level but also at the school principal level. On the one hand, pay for performance schemes (which generally reward gains in mean school level achievement) may induce administrators to just focus on efficiency matters. For instance, if teachers with stronger credentials are able to produce higher gains in test scores on high ability students than on low ability ones (while teachers with weaker credentials have no relative advantage), then school principals may be tempted to assign the best students with these high productive teachers in order to maximize mean school outcomes. On the other hand, programs such as the NCLB may operate in the opposite direction by giving school administrators incentives to concentrate on equity matters; and therefore allocating the best teachers to the less proficient students. In this sense, optimal sorting of teachers within schools can be improved after establishing the role of matching effects. The fact that previous empirical work indicate that most of the variation in teacher quality occurs within schools as opposed to between schools [Hanushek et al. (2005), Nye et al. (2004), and Koedel et al. (2007)] makes this point even more relevant. To sum up, interaction effects could constitute a fundamental aspect of policy design that deserves special attention.

Only a handful of papers have tried to tackle the role of student - teacher match effects; however,

these effects are obtained by interacting observed characteristics of students and teachers (e.g. race). Dee (2004), using random assignment data from the Tennessee STAR program, finds strong evidence for beneficial effects from matching the race of teacher and students; his results indicate that a year with an own-race teacher increased math and reading scores by roughly 2 to 4 percentile points. Clotfelter et al. (2006) argue that teachers with stronger credentials are more effective in raising achievement of the more advantaged students. Namely, they present evidence suggesting that math score returns to teacher experience (in North Carolina) are significantly larger for students that do not receive subsidized lunches. Hanushek et al. (2005) investigate matching effects by dividing students into three academic preparation classifications (based on initial scores) and compute the correlation between the teacher average gain for students in one category with the teacher average gain for students in the other categories. They find moderate positive correlations of 0.45 between the low and middle categories, 0.57 between the high and middle categories, and 0.31 between the low and high categories. In a similar vein, Koedel et al. (2007) test whether teacher effectiveness varies with initial student achievement by interacting individual teacher effects with a dummy variable that indicates if a student performed below the median in prior test scores. But they could not find evidence suggesting that teacher quality varies by student type. Finally, Lockwood et al. (2009) also estimate complementarity effects using past test scores, but in the context of a random effect model specification[7]. Their findings indicate that interactions account for about 10% of the total variation in teacher effects across all students.

This paper contributes to the analysis on student teacher complementarities by extending the existing literature on several dimensions. First, an analytical framework is developed with the aim to explicitly characterize the nature of student - teacher interactions. Second, the applied empirical strategy provides a novel opportunity to disentangle homogeneous educator effects from heterogeneous ones in a flexible way. Third, complementarity effects are analyzed by interacting student and teacher fixed effects instead of observed characteristics of teachers and students. Fourth, this study contributes to shed light on the mechanism behind test score production functions, showing this is crucial to understand how the different accountability policies operate.

Results, based on a North Carolina longitudinal database that links each student with their teachers during their schooling careers, indicate that interaction effects do matter. For example, one standard deviation increase in teachers slopes and intercepts for the 75^{th} percentile student in math (reading) leads to an increase of 0.237 (0.198) of a standard deviation in test scores where

[7] They estimate the model using Bayesian methods.

30% (32%) of that total effect corresponds to student - teacher interactions/complementarities. Moreover, an analysis of outcomes post NCLB indicates that teachers shifted their focus towards low performing students leading to a substantial improvement in test scores. For instance, around 60% of the total gains experienced by the bottom third of the student achievement distribution are due to changes in student - teacher interactions (i.e. slopes); suggesting that educators reallocated part of their attention on a given subgroup of pupils at the expense of others. In this regard, the top third of the student achievement distribution experienced no change in their test scores performance due to this adjustment in teachers' behavior, while the top quintile observes a decrease of -0.09 of a standard deviation. Finally, counterfactual exercises indicate that reallocation of teachers within school (conditional on grade) could not explain the observed change in students performance post NCLB.

The rest of the paper is organized as follows. Section 2 provides a brief description of the accountability programs in North Carolina. Section 3 presents a simple theoretical model. Section 4 describes the econometric strategy. Section 5 presents the data. Section 6 describes the results. Section 7 analyzes the effect of NCLB in teachers and principals behavior. Section 8 concludes.

2 Background on Accountability Programs

In the previous section, it was pointed out that teacher - student interaction effects mainly depend on the characteristics/preferences of the educator and the scheme of incentives in place (e.g. NCLB). Given that the analysis of this article is based on North Carolina data from elementary schools between the years 1997 and 2005, a brief description of the accountability programs that were in place during this period will help to characterize educators objective functions.

The schools in North Carolina have been subject to different accountability programs since the late 90's. In 1997, ABCs (Accountability for Basic skills and for local Control) was introduced with the aim to hold schools accountable for their value added. The main objective of this policy is to quantify how much children improve while being enrolled in a given institution. In this regard, teachers and staff in schools that are effective in rising student achievement receive salary bonuses[8]. These bonuses are based exclusively on the average gains in test scores for the cohort of students in the school during the year.

In the academic year 2002-2003, the No Child Left Behind program was layered on the top of the

[8] Bonuses range from $500 to $1500.

ABCs[9]. NCLB mandates that all students be proficient by 2014, and that each school must make Adequate Yearly Progress (AYP) towards meeting this objective, not only overall, but also for a set of demographic subgroups within in each school. Schools that fail to achieve the AYP standard for two consecutive years start to face sanctions, where their severity can increase depending on the past history. The first set of sanctions includes to offer students the possibility to transfer to higher-performing public schools in the same district. The second set imposes offering tutoring to selected students. Then, a "corrective action" is applied if the school continues failing. Basically, this could involve extending the school year or replacing some staff. The final stage in the sanction regime is restructuring, which could lead to turn over operation of the school to another entity and replace the staff, among other measures. Schools can exit the sanction regime by making AYP for two consecutive years[10].

The different designs of ABCs and NCLB suggest that teachers and administrators may have responded differently once each program took effect[11]. While ABCs may lead school principals and teachers to focus on efficiency goals rather than equity ones (in order to maximize school gains), the NCLB may work in the opposite direction (in order to improve lagging students performance). Therefore, two accountability programs with different scheme of incentives and rewards operated in North Carolina during the years that the data was collected.

3 Analytical Framework

In this section, a simple theoretical model is developed with the aim to characterize test scores production function and teachers behavior. The intention is to provide an analytic framework that gives structure to the mechanism that determines how complementarity effects operate, and how teachers respond to changes in average classroom characteristics. The solution to teachers maximization problem will lead to a set of final expressions that later will be estimated in the empirical part of the paper.

[9] The implementation of NCLB did not replace the ABCs program. On the contrary, both programs are still in place.

[10] See Ahn et al. (2009), Vigdor (2008) and Cooley et al. (2011) for a detailed description of this accountability program.

[11] In fact, Ladd et al. (2010) have shown (also based on North Carolina data) that educators respond to incentives and that the incentives to pay attention to students at different points of the achievement distribution differ between these two programs.

The test score production function for student i with teacher j in grade[12] g is assumed to be given by:

$$T_{ijg} = f(q_{1jg}, q_{2jg}, S_i, \varepsilon_{ijg}) \tag{1}$$

where q_{1jg} and q_{2jg} denote the qualities that characterize each teacher j in grade g on performing two different tasks, S_i represents the student fixed effect, and ε_{ijg} the idiosyncratic shock. Moreover, teacher abilities $\{q_{1jg}, q_{2jg}\}$ are defined as a function of the efforts exerted (i.e. e_{nj}) and the teachers' abilities (a_{nj}):

$$q_{1jg} = e_{1jg} + a_{1j} \tag{2}$$

$$q_{2jg} = e_{2jg} + a_{2j} \tag{3}$$

The presence of a vector of teacher qualities captures the notion that teachers are multi-tasking agents that perform complex jobs involving many duties [see Holmstrom and Milgrom (1991)]. More precisely, it is assumed that q_{1jg} has an homogeneous effect across all the students in the classroom, while q_{2jg} will depend on the specific interactions between the student and the teacher. In this regard, the functional form of the test scores production function is set as follows[13]:

$$T_{ijg} = q_{1jg} + S_i + q_{2jg}S_i + \varepsilon_{ijg} \tag{4}$$

The basic idea behind this representation is that each teacher can be characterized by an intercept and a slope (for a given pair $\{q_{1jg}, q_{2jg}\}$) where the term $q_{2jg}S_i$ will provide the necessary information to test the role of teacher - student complementarities[14] (see Figure 2 that illustrates this point).

[12] Grade and classroom are used as synonyms in this case.

[13] Neal (2011) also points out the importance of characterizing teachers as multi-task agents. In this regard, he considers the following human capital production technology: $h = f_1 t_1 + f_2 t_2 + e$, where t_1 and t_2 denote the time that the teacher devotes to two different tasks, f_1 and f_2 are constants, and e is a random shock.

[14] The implicit assumption that teacher inputs do not persist over time has also been assumed by Rockoff (2004) and Dee (2004). Kinsler (2011) has shown, using this same database, that this assumption leads to small bias in the teacher value-added variance (around 3% for reading exams). Moreover, not controlling for persistence understates the variance; therefore, in the worst case scenario, most likely will provide a lower bound of the true one. Jacob et al. (2009) find that teacher-induced learning has low persistence, with three-quarters or more fading out within one year. Finally, Lockwood et al. (2009) also point out that not including prior schooling inputs leads to quite small bias.

The following subsection discusses teachers decision process. More specifically, it presents the solution to their maximization problem and the theoretical predictions of the model.

3.1 Teachers Maximization Problem

Educators are expected to respond to the different scheme of incentives that were in place at a given point in time. Between the years 1997 and 2002, ABCs was the only accountability policy in effect in North Carolina. The fact that under this program rewards are a function of gains in test scores makes reasonable to assume that during these years, teachers' objective function involved the maximization of classroom average test scores[15]. Therefore, the maximization problem for teacher j in classroom g with N_g students is given by the following expression:

$$\max_{e_{1jg},e_{2jg}} U_{jg} = \max_{e_{1jg},e_{2jg}} \left[\gamma \frac{\sum_{i=1}^{N_g} T_{ijg}}{N_g} - C_{jg} \right] \tag{5}$$

where $\frac{\sum_{i=1}^{N_g} T_{ijg}}{N_g}$ represents course average test scores, C_{jg} the cost function, γ a weight parameter and $\{e_{1jg}, e_{2jg}\}$ the levels of effort that need to be chosen by each educator. Moreover, costs are assumed to have the following functional form:

$$C(e_{1jg}, e_{2jg}) = \frac{\alpha_1}{2}(e_{1jg})^2 + \frac{\alpha_2}{2}(e_{2jg})^2 \tag{6}$$

Substituting in for T_{ijg} and C_{jg} in (5) using (4) and (6), the teachers maximization problem becomes:

$$\max_{e_{1jg},e_{2jg}} U_{jg} = \max_{e_{1jg},e_{2jg}} \left[\gamma(a_{1j} + e_{1jg} + \overline{S_g} + (a_{2j} + e_{2jg})\overline{S_g} + \overline{\varepsilon_g}) - \frac{\alpha_1}{2}(e_{1jg})^2 - \frac{\alpha_2}{2}(e_{2jg})^2 \right] \tag{7}$$

where $\overline{S_g}$ denotes the average student fixed effect in the course. The first order conditions of this problem are given by:

$$\widetilde{e}_{1j} = \frac{\gamma}{\alpha_1} \tag{8}$$

$$\widetilde{e}_{2jg} = \frac{\gamma}{\alpha_2}\overline{S_g} \tag{9}$$

[15] Similar to Neal et al. (2010), the model does not intend to establish the socially optimal amount of effort per teacher or the socially optimal allocation of effort per student. Teachers are assumed to enjoy rents; and therefore the introduction of an accountability policy attempts to extract more effort of teachers.

On the one hand, equation (8) indicates that $\widetilde{e}_{1jg}$, which has an homogeneous effect for all the students in the classroom, depends on the "price" (α_1) of the effort, and the weighting parameter (γ). On the other hand, equation (9) shows that the slope that each teacher picks will depend on the course average characteristics. In this regard, the model is assuming that the mechanism in which peer effects operate is just through the election of $\widetilde{e}_{2jg}$ by the teacher.

After replacing teachers' optimal efforts ($\widetilde{e}_{1j}$, $\widetilde{e}_{2jg}$) in the test scores production function (equation 4), the following expression is recovered:

$$T_{ijg} = \underbrace{\frac{\gamma}{\alpha_1} + a_{1j}}_{q_{1j}} + \underbrace{\frac{\gamma}{\alpha_2}\overline{S_g}S_i + a_{2j}S_i}_{q_{2jg}S_i} + S_i + \varepsilon_{ijg} \tag{10}$$

Finally, after re-expressing $\frac{\gamma}{\alpha_1} + a_{1j} = q_{1j}$ and $\frac{\gamma}{\alpha_2} = \lambda$, the subsequent reduced form equation is obtained:

$$T_{ijg} = q_{1j} + \lambda\overline{S_g}S_i + a_{2j}S_i + S_i + \varepsilon_{ijg} \tag{11}$$

The previous expression will be estimated in the empirical section of the paper. In this regard, the aim is twofold. First, establish the dispersion of teachers intercepts and slopes to determine their relevance in test scores. Second, recover the magnitude of λ in order to test if teachers adjust their slopes based on average classroom characteristics.

The model implicitly assumes that teachers know the exact characteristics of each student (i.e. S_i). Alternatively, it can be replaced $\overline{S_g}$ with the classroom average test score in the previous year, $\overline{T}_{ijg-1}$. Including $\overline{T}_{ijg-1}$ instead of $\overline{S_g}$ implies that teachers do not know the effort exerted by students' previous teacher. Then, in addition to (11), it is also estimated the following equation that can also be derived from the model:

$$T_{ijg} = q_{1j} + \lambda\overline{T}_{ijg-1}S_i + a_{2j}S_i + S_i + \varepsilon_{ijg} \tag{12}$$

4 Econometric Strategy

Estimation of equation (12) is not trivial for two main reasons. First, identification and consistency properties of teacher estimators need to be study due to the fact that S_i are incidental parameters that may spoil the estimation of the parameters of interest. Second, estimation requires to implement an iterative algorithm given the large number of parameters to be obtained and the fact that within-group transformation fails to purge interactive fixed effects.

4.1 Identification and Consistency

Lee et al. (1993), Ahn et al. (2001), and Bai (2009) have studied the identification conditions of interactive fixed effect models. They show that additional restrictions to those used in traditional fixed effect estimations are required (in these models) in order to be identified. In this paper, teacher intercepts and slopes are normalized to have mean equal to zero. This normalization is suitable given that test scores will be standardized (i.e. mean zero and standard deviation one conditional on grade and year) when estimating equations (11) and (12).

This manuscript distinguishes from Lee et al. (1993), Ahn et al (1993) and Bai (2009) in several aspects. First, Lee et al. (1993) and Ahn et al. (2001) only consider the case where the interactive effect is the same for all individuals in the sample (i.e. $a_{2j} = a_2$), and second their models are focused on one component models[16]. Third, Bai (2009) works in the context of large T, while here $T = 3$ which is the maximum number of observations per student. In this regard, given that each student is observed from grade 3 to 5, the student fixed effect estimate will be unbiased, but inconsistent. This is a standard result in panel data models with large N fixed T, where N refers to the number of students. The fact that S_i are incidental parameters in the sense of Neyman and Scott (1948) (i.e. the number of S_i grows with sample size N) indicates that the consistency properties of teachers fixed effect estimators need to be analyzed, given that the incidental parameter (i.e. S_i) may spoil the estimation of the other parameters. However, theorem 1 shows that teacher effects parameters from model equation (12) are identified and consistent:

Theorem 1 *Let M denote the number of students who face at least two separate teachers and let T denote the number of observations for each of these M students. Then if:*

1. $E(\varepsilon_{i,j,g}, \varepsilon_{-i,-j,g}) = 0 \ \forall \ i \neq -i, \ j \neq -j, \ g \neq -g$
2. $E(\varepsilon_{ijg}) = 0 \ \forall \ i, j, g$
3. $E(\varepsilon_{ijg}^2) = \sigma^2 \ \forall \ i, j, g$
4. $\mathbf{q}^0 = \{q_{1j}^0, q_{2j}^0\}$ *lies in the interior of a compact parameter space* Θ

[16] For instance, Lee et al. (1993) consider the following specification: $y_{it} = X_{it}'\beta + \theta_t\delta_i + v_{it}$. Moreover, they reduce the analysis to the context of small T, because otherwise the specification "introduces too many parameters in the estimation problem," which might be problematic. However, the iterative algorithm developed next solves this problem.

Then, $\widehat{\mathbf{q}}$ *is identified and consistent for fixed* T *as* $M \to \infty$

See Appendix A for the proof of this theorem. Basically, the proofs of identification and consistency require to concentrate out the vector of unobserved students abilities (i.e. $\mathbf{S}$) of the original least squares problem. In this regard, consistent estimators can be obtained due to the fact that the estimated $\mathbf{S}$ are written as a function of teacher effects. For instance, if $\mathbf{S}$ were estimated in a first stage and later used in a second one to recover teacher slopes, then these parameters will (likely) be downward bias given that the problem of measurement error would hold.

The proof of identification and consistency for model equation (11) is more difficult given that it is not possible to concentrate out the vector of unobserved student abilities due to the presence of $\overline{S}$. However, Monte Carlo exercises indicate that it is possible to obtain tight estimates of $\{q_{1jg}, a_{2j}, \lambda\}$ around the truth[17].

4.2 Iterative Algorithm

Estimating this type of model is not simple for two main reasons. First, the within-group transformation cannot be implemented due to the presence of interactive fixed effects. Second, a substantial number of parameters must be recovered due to the large number of students and teachers in the sample[18]. Therefore, the econometric strategy requires an estimation procedure that includes several steps. In the spirit of Lee et al. (1991) and Arcidiacono et al. (2011), an iterative approach is implemented that toggles between estimating teacher fixed effects and student heterogeneity. In order to illustrate how the estimation works, the steps of the algorithm are presented next. First, consider the following model expression[19]:

$$T_{ijg} = q_{1j} + \lambda \overline{T}_{ijg-1} S_i + a_{2j} S_i + S_i + \varepsilon_{ijg} \tag{13}$$

where S_i represents student fixed effect, q_{1j} teacher intercept, a_{2j} determines (in part) the quality of the teacher-student match, $\overline{T}_{ijg-1}$ denotes the classroom average test scores in the previous period and ε_{ijg} represents the error term.

The algorithm begins with an initial guess of the parameters $S_i^{(0)}$. It then iterates on the following steps with the mth iteration:

[17] See Appendix C for Monte Carlo experiments based on model equation (11).

[18] Abowd and Kramarz (1999) recognize the fact that fixed effect models with large number of observations cannot be estimated directly. Instead, they consider a number of estimation techniques, none of which results in least squares.

[19] Appendix B describes the steps of the algorithm when $\overline{S}_g$ is included in the model equation instead of $\overline{T}_{ijg-1}$.

- **Step 1:** Using the initial guesses of the student fixed effects, calculate $Y_{ijg}^{(m)} = T_{ijg} - S_i^{(m)}$ and solve the least squares problem:

$$\{q_{1j}^{'(m)}, a_{2j}^{'(m)}, \lambda^{'(m)}\} = \arg \min_{e_{1j}, e_{2j}} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(Y_{ijg}^{(m)} - q_{1j} - S_i^{(m)} a_{2j} - \lambda \overline{T}_{ijg-1} S_i^{(m)} \right)^2 \tag{14}$$

- **Step 2:** Using $q_{1j}^{'(m)}$, $a_{2j}^{'(m)}$, and $\lambda^{'(m)}$ calculate $S_i^{(m+1)}$ based on the following expression ($j \in i$ denotes teachers of student i):

$$\frac{\sum_{j \in i} \left\{ \left[T_{ijg} - q_{1j}^{'(m)} \right] \left(1 + \lambda^{'(m)} \overline{T}_{ijg-1} + a_{2j}^{'(m)} \right) \right\}}{\sum_{j \in i} (1 + \lambda^{'(m)} \overline{T}_{ijg-1} + a_{2j}^{'(m)})^2} = S_i^{(m+1)} \tag{15}$$

where the previous expression avoids the minimization over all the S_i's. Equation (15) is obtained from the first order condition of the least squares problem when deriving with respect to S_i:

$$\begin{aligned} 0 &= \sum_{j \in i} \left(T_{ijg} - q_{1j} - S_i - S_i a_{2j} - \lambda \overline{T}_{ijg-1} S_i \right) \left(1 + \lambda \overline{T}_{ijg-1} + a_{2j} \right) \\ &= \sum_{j \in i} (T_{ijg} - q_{1j}) \left(1 + \lambda \overline{T}_{ijg-1} + a_{2j} \right) - S_i \sum_{j \in i} (1 + \lambda \overline{T}_{ijg-1} + a_{2j})^2 \end{aligned} \tag{16}$$

- **Step 3**: Repeat steps 1 and 2 until convergence of the parameters.

Similar iterative algorithms have also been used by Sargan (1964), Lee (1991), Lee at al. (1993), Brandt et al. (2009) and Arcidiacono et al. (2011). For instance, Sargan (1964) has shown that this type of iterative technique converge to parameter values which minimize the residual sum of squares[20].

4.3 Monte Carlo Evidence

Monte Carlo exercises were conducted in order to test how well the estimation strategy performs in small samples. The experiments were based on equation (12)[21]. The aim is to show that the algorithm is capable of providing tight estimates of $\{q_{1jg}, a_{2j}, \lambda\}$ around the truth.

[20] As with full solution methods local minima is possible. In order to check for multiple optima, different starting values were used in the estimation.

[21] See Appendix C for Monte Carlo experiments based on equation (11).

The structure of the data was chosen to mimic the aspects of the North Carolina primary school characteristics. In this regard, three cohorts of 25 students were created (which is the average size of an elementary classroom in North Carolina), where each student is observed three times. Students fixed effects were assumed to have mean -1 and variance 1 for cohort 1, mean 0 and variance 1 for cohort 2 and mean 1 and variance 1 for cohort 3. Moreover, students are assumed to have different classmates during their schooling years. This implies that students from different cohorts can share the same classroom. Finally, three teachers were created with intercepts equal to -1, -0.5, 1.5 and slopes equal to -0.5, 0.15, and 0.35 respectively. Each of the teachers has a different student composition of their class in each year but they can have twice the same student. Finally, teachers were sorted into the different classrooms randomly. Test scores measurement error are assumed independent and identically distributed across grades according to $N(0, \sigma^2)$. The values of σ were chosen based on what is observed from the data. Finally, λ was set to be equal 0.5. Table 1 presents the mean of the recovered parameters after estimating 100 times the model equation for each value of σ. Results indicate that the algorithm preforms quite well, providing tight estimates of the true parameters and centered around the truth given the small sample size.

5 Data

The data comes from administrative records maintained by the North Carolina Education Research Data Center (NCERDC). This longitudinal database contains yearly test scores for each student in mathematics and reading in elementary,[22] middle, and high school. The sample used in this paper covers the period 1997-2005 and it only includes grades 3 to 5[23]. Encrypted identifiers make it possible to track students over their educational careers. As well as linking students to their teacher[24] and school in each year.

NCERDC records also include extensive information on students, teachers and schools characteristics. Data on students include parents education, ethnicity, gender, exceptionality classifica-

[22] More specifically, from grade 3 and above. Students in grades 1 and 2 do not have to take end of grade tests, but at the beggining of grade 3 they do have to take an exam.

[23] In upper grades, students can take courses with other than their full-time assigned teacher, making the estimation of teacher fixed effects problematic.

[24] The North Carolina data does not indentify students' teachers directly, but they do identify the person who administered the end of grade tests. In elementary grades, this was usually the regular teacher. In order to minimize this inconvenience, similar strategies to those in Clotfelter et al (2006) and Rothstein (2009) were applied.

Monte Carlo Evidence							
Model: $Test_{ij} = q_{1j} + a_{2j}S_i + \lambda\overline{T}_{ijg-1}S_i + S_i + \varepsilon_{ijg}$							
	Teacher Intercepts			Teacher Slopes			Slope Class. Adj.
True Parameters	q_{11}= -1	q_{12}= -0.5	q_{13}= 1.5	a_{21}= -0.5	a_{22}= 0.15	a_{23}= 0.35	$\lambda = 0.5$
Estimates	-0.999	-0.503	1.502	-0.502	0.151	0.351	0.500
(σ_ε= 0.25)	(0.024)	(0.026)	(0.032)	(0.024)	(0.031)	(0.037)	(0.032)
Estimates	-0.997	-0.498	1.494	-0.501	0.141	0.359	0.492
(σ_ε= 0.3)	(0.032)	(0.027)	(0.043)	(0.025)	(0.036)	(0.040)	(0.036)
Estimates	-0.992	-0.501	1.493	-0.498	0.149	0.349	0.502
(σ_ε= 0.5)	(0.047)	(0.053)	(0.063)	(0.051)	(0.058)	(0.077)	(0.061)

Table 1: Monte Carlo Evidence for different standard deviation in test scores measurement error. The reported parameters were obtained after estimating the model 100 times for each value σ_ε, and then taking the average conditional on σ_ε. In parenthesis, it is presented the standard errors of the coefficients.

tions, participation in the federal free and reduced price lunch subsidy program. On the teacher side, the information involves final educational attainment, experience and the score on the test used to obtain a teaching license, among other variables. Finally, schools characteristics contain name, district, demographic characteristics of the student and teacher body, and school overall performance in fulfilling the NCLB requirements.

As Clotfelter et al. (2006) point out, North Carolina is an appropriate state for teacher effectiveness analysis because its exams are closely aligned with what students are expected to know[25]. Therefore, test scores are likely to measure more fully what teachers have taught than in many other states. Moreover, the state is relatively large and exhibits substantial variation across schools in terms of the racial and socioeconomic mix of the students.

The analysis focuses on the five biggest counties in North Carolina (i.e. Cumberland, Wake, Forsyth, Guildford and Charlotte-Mecklenburg), representing around 30% of the total student population between grades 3 and 5[26]. Table 2 shows the evolution of math and reading average test scores (for selected years) that were standardized with respect to the mean and standard deviation of year 2001 conditional on grade[27]. Two main conclusions can be obtained from this table. First, there is a positive trend in mean test scores indicating an improvement in students performance. Second, there is also a substantial reduction in the standard deviation, denoting a compression of the test scores distribution. In section 7, it is shown that these empirical regularities can be explained in part by a higher increase in the performance of lagging students relative to high performing students due to a change in teacher - student interaction effects after NCLB.

Teachers with less than 20 observations across the sample period as well as classrooms with less than 10 students were not included in order to minimize problems of measurement error when estimating fixed effects. Moreover, classrooms with more than 35 students were also eliminated due to possible data miscoding. The total number of student-year observations for the period 1997-2005 is more than 370000, while the total number of teachers included is more than 5200.

[25] The scores are comparable across time and grades through the use of a developmental scale. The developmental scale is created from the number of correctly answered questions on the standardized test. Each point of the developmental scale measures the same amount of learning. For instance, a child who shows identical growth on this scale in two consecutive grades is interpreted as learning equal amounts in each year.

[26] Appendix D shows that the mean test scores in these 5 counties are similar to those observed in the full sample.

[27] Test scores experienced a change of scale in year 2000 for math and 2002 for reading; however a table that matches old scores with the new ones makes it possible to perform comparisons across years.

Evolution of Average Math and Reading Test Scores for Selected Years						
	1998	2001	2004	1998	2001	2004
		Math			Reading	
Grade 3	-0.162	0	0.305	-0.151	0	0.141
	(0.995)	(1)	(0.799)	(1.012)	(1)	(0.917)
Grade 4	-0.081	0	0.359	-0.057	0	0.188
	(1.140)	(1)	(0.900)	(0.980)	(1)	(0.901)
Grade 5	-0.248	0	0.265	-0.161	0	0.141
	(1.072)	(1)	(0.909)	(1.055)	(1)	(0.859)

Table 2: Mean and standard deviation of math and reading tests scores for selected years only considering the five biggest counties in North Carolina (i.e. Cumberland, Wake, Forsyth, Guildford and Charlotte-Mecklenburg). Scores were standardized with respect the mean and standard deviation of year 2001.

6 Results

6.1 Teachers Efficacy: Homogeneous and Complementarity Effects

This section provides a first set of results that disentangle the importance of homogeneous from heterogeneous teacher effects. Estimation outcomes were obtained separately for math and reading exams for the period 1997 - 2002 (years in which only the ABCs program was in place). Given that there are almost no differences in the results when estimating model equations (11) or (12), only the outcomes from the following test scores production function are reported in this section[28]:

$$T_{ijg} = q_{1jg} + \lambda \overline{T}_{ijg-1} S_i + a_{2jg} S_i + S_i + \beta I(\exp r < 1yr) + \varepsilon_{ijg} \tag{17}$$

where $I(\exp r < 1yr)$ denotes a dummy variable indicating whether a teacher has less than 1 year of experience. Test scores were standardized to have zero mean and standard deviation one, conditional on year and grade, in order to control for trends in test scores.

Table 3 presents the standard deviation and correlation of teachers intercepts ($\widehat{q_{1jg}}$) and slopes ($\widehat{a_{2j}}$) with the aim to provide an initial examination of the estimation outcomes. First, it is important to highlight that a likelihood ratio test for each type of exam (i.e. math and reading) rejects the null hypothesis that teacher slopes are jointly equal to zero; this indicates that complementarity

[28] Appendix E reports estimation results based on model equation (11). To compare results across models look at Tables 3 and E1.

	Estimation Results			
	Math		Reading	
	Intercept	Slope	Intercept	Slope
Std. Dev.	0.204	0.161	0.153	0.127
$corr(Int_{jMath}, Int_{j\,\mathrm{Re}\,ad})$		0.455		
$corr(Slope_{jMath}, Slope_{j\,\mathrm{Re}\,ad})$		0.090		
$corr(Int_{jMath}, Slope_{jMath})$		0.269		
$corr(Int_{j\,\mathrm{Re}\,ad}, Slope_{j\,\mathrm{Re}\,ad})$		-0.067		
λ	0.006		0.011	
	(0.004)		(0.005)	
$\beta * I(\exp r < 1yr)$	-0.055		-0.033	
	(0.007)		(0.008)	
L-R test (P-Value)	0.001		0.001	

Table 3: Standard deviation and correlation of teacher fixed effects (i.e. intercepts and slopes) recovered from the estimation of equation (17) on different subjects. Likelihood ratio (L-R) tests results that analyze the joint significance of teachers slopes are also reported. λ indicates how teachers adjust their slopes based on classroom characteristics and β denotes the effect of having a teacher with less than one year of experience.

effects do matter at least just from an statistical point of view. Second, the estimates show that one standard deviation increase in teacher intercepts for math and reading exams lead to an increase in students test scores of 0.204 and 0.153 of a standard deviation respectively. In this regard, Kinsler (2011), using this same database, but estimating a different model, finds similar results (i.e. 0.25 for math and 0.14 for reading) for these parameters. Third, the dispersion of teachers slopes are also large for both subjects; being around 80% of the intercepts' standard deviation.

Next, correlation results are presented with the aim to provide a snapshot on teachers performance across subjects and tasks. Table 3 shows that the correlation between teachers intercepts (slopes) across disciplines is 0.455 (0.090); suggesting heterogeneous teacher effectiveness across subjects[29]. Lastly, the correlation between slopes and intercepts for a given subject shows that this relationship is statistically far from been equal to 1, implying that the example provided in Figure 2 is actually frequent in the data.

[29] An alternative explanation might be sampling error, an issue that is analyzed in section 6.3.

6.2 Do Teachers Adjust Their Slopes Based on Mean Classroom Characteristics?

The aim of this subsection is to study whether or not teachers adjust their slopes based on classroom characteristics. This type of adjustment could be interpreted as educators trying to modify the structure of the course in order to make it more suitable for the average characteristics of the students in the classroom. The parameter λ from equation (17) is the one that provides information on this regard. Results for math and reading exams indicate that $\lambda_{Math} = 0.006$ and $\lambda_{\text{Re}ad} = 0.011$[30]. Only $\lambda_{\text{Re}ad}$ is statistically significant different from zero; however, its effect in terms of students test scores is almost null. For instance, one standard deviation increase in $\overline{T}_{ijg-1}S_i$ for reading exams leads to an increase in 0.5% of a standard deviation in test scores. This null effect could be explained by the fact that teachers face (in general) low variability in mean classroom characteristics across academic years. Therefore, they may have no incentives to incur in slopes adjustments, when the cost of doing so is relatively high. However, as it is shown later, the implementation of NCLB has created enough changes to the scheme of incentives such that teachers adjusted their slopes with the aim to improve the performance of a given subgroup of students in the classroom.

6.3 Sampling Error

The fact that teacher fixed effects constitute noisy measures of true teacher value added may have implications for the estimates of teachers quality dispersion; leading to inflate standard deviations. More specifically, if each fixed effect coefficient is comprised of two components - the true signal of teacher quality and the sampling error - then the recovered variance of q_{nj} (where $n = 1$ *or* 2) is given by:

$$Var(\widehat{q_{nj}}) = Var(q_{nj}) + Var(\xi) \tag{18}$$

where ξ denotes the sampling error and $cov(q_{nj}, \xi)$ is assumed to be zero[31]. However, it is expected that the size of $Var(\xi)$ will decrease substantially if the analysis is performed on a subsample of teachers with a large number of observations. A drawback of following this path is that the variability in teachers quality could also decrease for other reasons different to sampling error. For

[30] Results based on model equation (11), indicate that $\lambda_{Math} = 0.022$ and $\lambda_{\text{Re}ad} = 0.0001$, with only λ_{Math} being statistically significant different from 0.

[31] Similar assumptions have been made in Aaronson et al (2007) and Kinsler (2011), among others.

Standard. Deviation. of Teachers Fixed Effect Cond. on Number of Classrooms				
Classrooms Per Teacher	Intercept$_{math}$	Slope$_{math}$	Intercept$_{read}$	Slope$_{read}$
2	0.20	0.16	0.17	0.15
3	0.20	0.13	0.16	0.13
4	0.19	0.13	0.15	0.11
5	0.17	0.11	0.12	0.09
6	0.17	0.11	0.13	0.09

Table 4: Standard deviation of teachers fixed effects for math and reading conditional on the number of classrooms per teacher.

instance, teachers with high number of observations are more experienced or could be more similar to each other due to, for example, a selection process, then part of the true dispersion in teacher effectiveness would be lost. Therefore, it is reasonable to claim that the standard deviation of teacher fixed effects conditional on those with highest number of observations likely provide a lower bound of the true one.

Table 4 shows the dispersion of teachers intercepts and slopes conditional on the number of courses. Results indicate that as the number of classrooms per educator increase the standard deviations of intercepts and slopes decrease for both subjects (i.e. math and reading). However, after 5 courses per teacher these values become quite stable and remain quite constant. In this regard, the last row of Table 4 shows that likely lower bounds for the variability in teachers intercepts (slopes) are 0.17 (0.11) for math and 0.13 (0.09) for reading.

The correlation coefficients presented in Table 3 define, for example, the relationship between $(q_{n,j,subject} + \xi_{n,j,subject})$ and $(q_{n,j,-subject} + \xi_{x,j,-subject})$. In this sense, sampling error may lead to underestimate correlation effects. But if the correlation of true teacher quality across subjects for all teachers is assumed to be the same then, following Rockoff (2004), it is possible to recover the direction of the bias introduced by the sampling error in the estimated teacher fixed effects. Given that the error is smaller for teachers with a greater number of student-year observations, then it is possible (again) to calculate the correlation coefficient between $q_{x,j,\widehat{subject}}$ and $q_{x,j,\widehat{-subject}}$ for a subset of teachers who have a relatively high number of students to that of the entire teacher sample. Results indicate that sampling error is biasing correlations of teacher intercepts and slopes across subjects toward zero. For instance, Table 5 shows that the correlation of teachers intercepts

Corr. of Fixed Effects Cond. on Obs.	
$corr(q_{1jMath}, q_{1j\,\mathrm{Re}\,ad})$	0.563
$corr(a_{2jMath}, a_{2j\,\mathrm{Re}\,ad})$	0.232

Table 5: Correlation of teachers fixed effect (i.e. intercepts $\{q_{1jMath}, q_{1j\,\mathrm{Re}\,ad}\}$ and slopes $\{a_{2jMath}, a_{2j\,\mathrm{Re}\,ad}\}$) conditional on at least 60 observations per teacher.

(slopes) is 0.563 (0.232), when considering teachers with more than 60 observations[32], which is 0.108 (0.142) points higher than the one showed in Table 3. Finally, it is important to mention that the sign of the biases are consistent with the findings of Rockoff (2004) and Koedel (2007).

To conclude, sampling errors seem to artificially inflate and deflate the true standard deviations and correlations respectively. However, none of the conclusions derived in subsection 6.1 have lost its economic relevance due to adjustments for sampling error. These corrections most likely provide lower bounds (when analyze standard deviations) of the true effects.

6.4 Quantifying Teachers Effectiveness

In order to analyze in more detail the relevance of teacher effectiveness, Table 6 shows simulated changes in test scores for different percentiles of student ability, using both the standard deviation obtained directly from the estimation outcomes (see Table 3) and the likely lower bounds presented in the last row of Table 4. This simple exercise measures what would happen to test score performance if a given student experiences a 1 standard deviation increase in his/her teacher characteristics (i.e. intercept and slope, but once at a time) relative to being taught by a teacher with intercept and slope equal 0 (i.e. the mean teacher). Results in panel A[33] of Table 6 show that for the 75^{th} percentile student in math or reading, an increase in 1 sd in the slope of his/her teacher produces an increase in scores which is around half of a similar increase in the teacher intercept[34]. On the contrary, low ability students (e.g. 25^{th} percentile) that experienced a similar increase in the teacher slope, suffer a decrease in tests scores relative to having the "mean slope teacher". This outcome can be characterized, for example, by a situation where the teacher is focusing on improving the performance of the best students at the expense of harming the learning

[32] This correlation coefficient does not change when consider a subgroup of teachers with more observations.

[33] The standard deviations used in this case are the ones that were obtained directly from the data (see Table 3).

[34] The increase in math (reading) test scores due to a change in the slope is 0.107 (0.087) and due to a change in the intercept is 0.204 (0.153).

	Quantifying Teacher Effectiveness					
	Math			Reading		
	Panel A: Std. Dev. not corrected for sampling error					
	SP 25^{th}	SP 50^{th}	SP 75^{th}	SP 25^{th}	SP 50^{th}	SP 75^{th}
Mean Teacher Inter. + 1 SD	0.204	0.204	0.204	0.153	0.153	0.153
Mean Teacher Slope + 1 SD	-0.113	-0.003	0.107	-0.088	0.006	0.087
Total	0.091	0.201	0.311	0.065	0.159	0.240
	Panel B: Std. Dev. corrected for sampling error					
Mean Teacher Inter. + 1 SD	0.165	0.165	0.165	0.134	0.134	0.134
Mean Teacher Slope + 1 SD	-0.077	-0.002	0.072	-0.063	0.004	0.064
Total	0.088	0.163	0.237	0.071	0.138	0.198

Table 6: Change in test scores for different percentiles of student ability (SP) due to a 1 standard deviation increase in teacher intercept or slope (once at a time). Total denotes the sum of both changes.

process of the low ability students. Finally, panel B shows the results of this same exercise but this time using the likely lower bound standard deviations recovered from the last row of Table 4. The results indicate, as expected, lower changes in scores; however the ratio of the teacher slope to teacher intercept remains relatively constant. To sum up, teacher - student interactions constitute an important component of total teacher effectiveness. In this regard, one standard deviation increase in teachers slopes and intercepts for the 75^{th} percentile student in math (reading) leads to an increase of 0.237 (0.198) of a standard deviation in test scores, being 30% (32%) of that total effect due to complementarity effects.

As a final robustness check, it is important to analyze the variability in student and teacher characteristics within schools. For example, if student ability and teacher slopes were homogeneous then the role of complementarities would not be quite relevant to explain variability in students performance in a given institution. In this regard, Table 7 shows that the average standard deviation of test scores within school is 0.93 for math and reading exams; indicating a substantial dispersion in students performance. In terms of teacher characteristics, Table 7 indicates that the average standard deviation of teachers intercepts (slopes) within school is around 20% (30%) smaller than the dispersion registered in the total sample. Moreover, the ratio standard deviation teacher slope - teacher intercept at the total sample and the school level is similar. Finally, the average dispersion

of intercepts (slopes) within school conditional on grade[35] represents 70% (60%) of the total sample dispersion. To sum up, the evidence indicates that substantial variability in students and teachers characteristics can also be found within school.

Standard Deviation of Teacher Intercepts and Slopes				
	Reading		Math	
	Intercept	Slope	Intercept	Slope
Std. Dev. in the Total Sample	0.15	0.13	0.20	0.16
Average Std. Dev. within School	0.12	0.10	0.16	0.11
Average Std. Dev. within School Cond. on Grade	0.11	0.08	0.14	0.09
Average Std. Dev. Test Scores within School	0.93		0.93	

Table 7: Standard deviation of students and teachers characteristics based on: the total sample, within school, and within school conditional on grade (i.e. only the schools with more that one course per grade were considered in the last case).

7 NCLB Program

In Section 2, it was mentioned that No Child Left Behind was layered on the top of the ABCs[36] in the academic year 2002-2003. Therefore, it is expected that teachers adjusted their behavior in response to this new policy. More specifically, the fact that teachers are multi-task agents indicates that a change in the scheme of incentives will affect how educators allocate efforts among students at different points of the achievement distribution. In this regard, the findings in the literature show that this program did change the shape of the distribution of student achievement as measured by scores on particular tests. On the one hand, Neal et al. (2010) indicate that the combination of No Child Left Behind and similar reforms introduced in Chicago resulted in an improvement in scores on the state-mandated reading and mathematics tests for students in the middle of the achievement distribution, but not for students at the bottom of the distribution and not consistently for students at the top. On the other hand, Ladd et al. (2009) find, using

[35] Schools with more that one course per grade were considered in this case.

[36] The implementation of NCLB did not replace the ABCs program. On the contrary, both programs are still in place.

the North Carolina database[37], that NCLB resulted in an increase in the performance of lagging students (both those near the proficiency cut score and those well below it) at the expense of the achievement of high-performing ones. In this sense, Murnane (2011) suggests that the difference in findings may arise, because the proficiency standard in North Carolina is relatively low. Reback (2008) shows, based on data from the Texas Education Agency, that schools responded to NCLB by taking broad measures that help all low achieving students, while for high achieving ones this was not the case.

Prior evidence suggests that NCLB has influenced teachers behavior, leading them to focus on less advantaged students at the expense of high ability ones. Here, it is shown what share (if any) of this increase in test scores was due to teachers exerting more effort overall (i.e. increase their intercepts) and what share can be attributed to changes in teacher - student interactions (i.e. variation in teachers slopes). The findings in the literature and the design characteristics of the program indicate that most likely educators decrease their slopes (by giving more weight to lagging students), and/or increase their intercepts once NCLB took effect. A simple extension of the theoretical model presented in Section 3, makes it possible to test this likely change in teachers behavior. Given that allowing for adjustments in intercepts and slopes at the teacher level (post NCLB) involves estimating four fixed effects per teacher (which requires a substantial amount of observations per teacher), a simpler model is implemented under the assumption that changes in teacher intercepts and slopes post NCLB occurred at the school level. Therefore, the teacher's maximization problem can be written as follows:

$$\begin{aligned}\max_{e_{1jg},e_{2jg}} U_{jgs} &= \max_{e_{1jg},e_{2jg}} \gamma((a_{1jg}+e_{1jg})(1+\zeta_{1s}I(NCLB))+\overline{S_g} \\ &\quad +(a_{2jg}+e_{2jg})(\overline{S_g}+\zeta_{2s}I(NCLB))+\overline{\varepsilon_g})-\frac{\alpha_1}{2}(e_{1jg})^2-\frac{\alpha_2}{2}(e_{2jg})^2 \qquad (19)\end{aligned}$$

The previous expression is similar to equation (7) with the only difference that indicators for whether NCLB was in place or not were included (i.e. $I(NCLB)$); subscript s denotes that teacher adjustments after NCLB are made at the school level. A priori, it is expected $\zeta_{1s} > 0$, implying that teachers increased their intercepts (which will benefit all the students in a classroom) and $\zeta_{2s} < 0$, meaning that teachers tried to target less advantage students in a given course instead of the average one. Equation (19) could be understood as teachers trying to maximize a weighted mean of classroom test scores (i.e. $\frac{\sum_{i=1}^{N_g} w_i T_{ijgs}}{N_g}$) instead of just the average (i.e. $\frac{\sum_{i=1}^{N_g} T_{ijgs}}{N_g}$), where w_i

[37] This same database is used in this paper.

is expected to be higher for low performing students, which it will be reflected by the sign of the $\zeta's$ coefficients[38].

The first order conditions of the maximization problem are:

$$\overline{e}_{1jgs} = \frac{\gamma}{\alpha_1} + \frac{\nu_s * I(NCLB)}{\alpha_1} \tag{20}$$

$$\overline{e}_{2jgs} = \frac{\gamma}{\alpha_2}\overline{S_g} + \frac{\varpi_s * I(NCLB)}{\alpha_2} \tag{21}$$

where $\nu_s = \gamma\zeta_{1s}$ and $\varpi_s = \gamma\zeta_{2s}$. After replacing $\overline{e}_{1jgs}$ and $\overline{e}_{2jgs}$ in test scores production function (i.e. equation 4), the following expression is recovered:

$$T_{ijg} = \underbrace{\frac{\gamma}{\alpha_1} + \frac{\nu_s}{\alpha_1}I(NCLB) + a_{1j}}_{q_{1jgs}} + \underbrace{\frac{\gamma}{\alpha_2}S_i\overline{S_g} + \frac{\varpi_s}{\alpha_2}S_iI(NCLB) + a_{2j}S_i}_{q_{2jgs}S_i} + S_i + \varepsilon_{ijgs} \tag{22}$$

Denoting $\frac{\gamma}{\alpha_1} + a_{1j} = \overleftrightarrow{q}_{1jg}$, $\frac{\nu_s}{\alpha_1} = \rho_s$, $\frac{\gamma}{\alpha_2} = \pi$ and $\frac{\varpi_s}{\alpha_2} = \varphi_s$, leads to the following reduced form equation, which will be estimated in the empirical part of the paper:

$$T_{ijg} = \underbrace{\overleftrightarrow{q}_{1jg} + \rho_s * I(NCLB)}_{q_{1jgs}} + \underbrace{\pi\overline{S_g}S_i + \varphi_sS_iI(NCLB) + a_{2j}S_i}_{q_{2jgs}S_i} + S_i + \varepsilon_{ijgs} \tag{23}$$

where (as it was discussed above) it is expected that $\rho_s > 0$ and $\varphi_s < 0$, given that γ, α_1 and α_2 have to be positive. Due to the same reasons as in section 3.1, a slightly different version of equation (23), where $\overline{S_g}$ is replaced by $\overline{T}_{ijg-1}$, is also estimated.

7.1 Preliminary Evidence and Results

In order to provide a preliminary picture of changes in students performance before and after NCLB in North Carolina, the evolution of raw math scores is analyzed. In this regard, Table 8 shows standardized tests scores with respect to year 2001 in grade 3 in North Carolina for different students percentiles pre (2001-2002) and post (2003-2004) NCLB[39]. First, it is important

[38] Reback et al. (2011) find that after the implementation of NCLB teachers report greater concern over how student test performance will affect their job security and that untenured teachers in high-stakes grades/subjects work longer hours. In a similar vein, Murnane et al. (2010) argue that NCLB may threaten the jobs of those teachers that work in low performing schools. In this regard, state policy dictates that teachers in North Carolina receive tenure after teaching in the state's public schools for four consecutive years.

[39] See table F1 of Appendix F for a similar table that includes grades 4 and 5. In addition, Figure F1 shows the distribution of raw grades in math.

Evolution of Math Test Scores by Year: Grade 3				
Student Percentile	2001	2002	2003	2004
10^{th}	-1.24	-1.24	-0.76	-0.76
25^{th}	-0.76	-0.52	-0.15	-0.15
50^{th}	-0.03	0.09	0.33	0.33
75^{th}	0.69	0.81	0.81	0.81
85^{th}	1.05	1.17	1.17	1.17
90^{th}	1.29	1.41	1.29	1.29
95^{th}	1.65	1.77	1.53	1.53
Mean	0	0.11	0.32	0.31
Std. Dev.	1	0.99	0.81	0.79
Observations	28726	28670	29334	29543

Table 8: Evolution of math test scores pre (2001-2002) and post (2003-2004) NCLB for different percentiles of student achievement in grade 3. Test scores were standardized with respect to the mean and standard deviation of year 2001.

to notice that there was a substantial decrease in test scores dispersion after the implementation of this accountability program. More specifically, the standard deviation dropped from 0.99 to 0.81 between 2002 and 2003. Second, standardized test scores for the 25^{th} percentile student and below experienced an increased of 0.48 points during this same period of time, while for the 75^{th} did not change[40]; suggesting that students at the bottom of the achievement distribution have shown a higher improvement in their performance after the implementation of NCLB, than their high ability counterparts[41].

The data use in the estimation covers the period the period 2000 - 2005 (i.e. three years before and three years after NCLB took effect). Test scores were normalized with respect to the mean and the standard deviation of year 2001 conditional on grade. As a consequence, a linear trend was included in order to control for grade inflation. Only math test scores were used for the

[40] A two-sample Kolmogorov-Smirnov test for equality of distribution functions for the years 2002-2003 rejects the null hypothesis of same distribution function. In a similar vein, a test of difference in means reject the null hypothesis of equal means.

[41] For these particular years, none sanctions were applied given that these were the first two years that the program was in effect.

estimation given that reading score exams suffered a change of scale the same year that NCLB was implemented. This sample includes around 290 schools[42]. Finally, due to the fact that the results obtained from estimating the models with $\overline{T}_{ijg-1}$ or $\overline{S_g}$ are quite similar, only one of these models are discussed. Then, the estimated model equation is the following one:

$$T_{ijg} = \underbrace{\overleftrightarrow{q}_{1jg} + \rho_s * I(NCLB)}_{q_{1jgs}} + \underbrace{\pi\overline{T}_{ijg-1}S_i + \varphi_s S_i I(NCLB) + a_{2j}S_i}_{q_{2jgs}S_i} + S_i + \delta trend + \beta I(\exp r < 1yr) + \varepsilon_{ijgs} \tag{24}$$

where *trend* denotes a linear trend and $\exp r$ denotes a dummy variable indicating whether a teacher has less than 1 year of experience.

Estimation Results Equation (24)				
$Mean(\rho_s)$	$Mean(\varphi_s)$	δ	π	β
0.178	-0.209	0.108	0.018	-0.056
		(0.001)	(0.003)	(0.006)

Table 9: Estimation results of model equation 24, standard errors in parenthesis. $Mean(\rho_s)$ and $Mean(\varphi_s)$ denotes the mean of these parameters across schools

Table 9 shows that after the implementation of NCLB, all students experienced on average an increase in test scores of 0.178 of a standard deviation which can be interpreted (based on the model) as a change in intercepts (i.e. ρ_s). But teacher slopes also changed. The decrease in slopes (i.e. φ_s) benefits less advantaged students and hurts most advantaged ones. For instance, the bottom third of the student achievement distribution experienced a gain of 0.464 of a standard deviation (once NCLB took effect), where 60% of this increase can be explained by changes in student - teacher interactions[43]. This outcome suggests that educators have reallocated part of their attention on less advantaged students at the expense of the most proficient ones. For example, the top third of the achievement distribution experienced no gains (-0.017), while the top quintile shows a decrease of -0.09 of a standard deviation. Therefore, these results shed light on the mechanism behind the outcome of NCLB (i.e. teachers increasing their intercepts but also reducing substantially their slopes).

[42] Table F2 in Appendix F shows that students characteristics remained constant pre and post NCLB. Table F3 shows a similar pattern for teachers.

[43] See table G1 of appendix G for estimation results were ρ and φ are not allowed to change by school. Table G2 presents results when $\overline{T}_{ijg-1}$ is replaced by $\overline{S}_g$,

Figure 4 shows the distribution of the change in intercepts (i.e. ρ_s) and slopes (i.e. φ_s) across schools after NCLB. In general, most institutions have responded to this policy similarly, i.e. increasing their intercept but decreasing their slopes[44].

In order to analyze how these changes are correlated with schools characteristics, figures 5 and 6 show the change in intercepts and slopes in each school conditional on certain school covariates pre NCLB. First, Figures 5A and 5B show that the institutions with highest mean test scores pre NCLB increase their intercepts less but decrease their slopes more[45]. Second, Figures 6A and 6B indicate that the schools that changed the most their intercepts (slopes) were the ones with lowest (highest) mean intercept (slope) pre NCLB. This implies that schools have adjusted on those margins where they have more room to do so (e.g. flat slope schools have less room to set flatter slopes, therefore they have to shift their intercepts). In order to analyze this point in more detail, Table 10 shows the average intercept and slope pre NCLB within school conditional on the change in intercepts and slopes post NCLB. This table indicates that the schools that show an increase above the median in their intercepts also show lower mean intercepts pre NCLB (i.e. -0.013 vs. 0.042). Similarly, the schools that decreased their slopes the most where the ones with higher mean slopes pre NCLB (i.e. 0.026 vs. -0.023).

Mean Intercept and Slope Pre NCLB Conditional on Change in Slope and Intercept Post NCLB		
	Mean Intercept	Mean Slope
Above Median (increase) Intercept Change	-0.013	-0.005
Below Median (increase) Intercept Change	0.042	0.001
Above Median (decrease) Slope Change	0.033	0.026
Below Median (decrease) Slope Change	0.003	-0.023

Table 10: Mean school intercept and slope pre NCLB conditional on the changes observed on these variables post NCLB.

[44] The standard deviations of both kernel distributions are around 0.1

[45] A small fraction of students changed school due to NCLB sanctions. Cooley et al. (2011) indicate, using this same database, that the effects of sanctions on school choice are quite small in magnitude and fairly comparable across sanction statuses. With 10 choices in the district, students in sanctioned schools are about 1% more likely to change schools. With 20 choices, they are about 2% more likely to change schools.

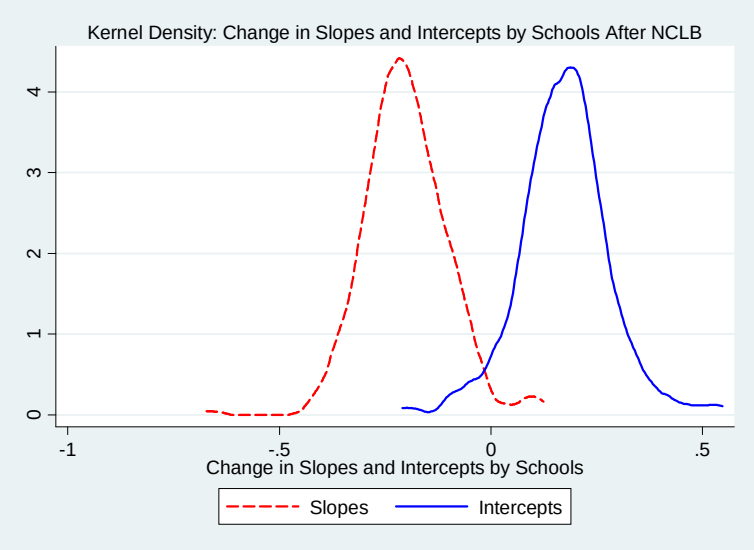


Figure 4: Distribution of the change in slopes and intercepts across schools post NCLB

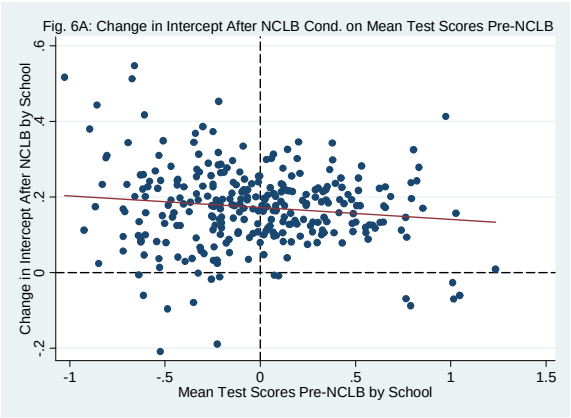
7.2 Relevance of Teachers Reallocation After NCLB

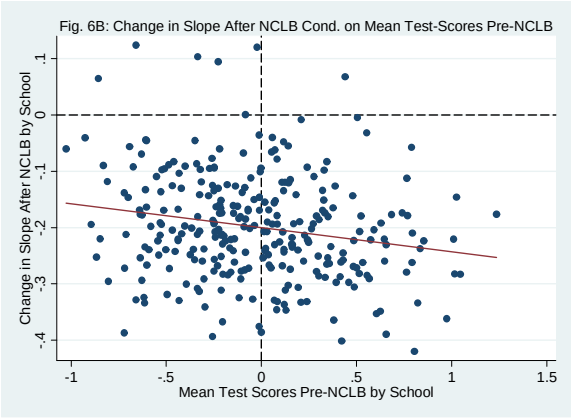
The implementation of NCLB may have lead school principals to reallocate teachers within schools. In order to analyze the relevance of this mechanism, two exercises were performed. First, predicted test scores were calculated for the year 2003 after turning off the parameters of NCLB. A priori, it is expected that if reallocation of teachers was substantial, then differences in performance should be registered between the predicted scores (without NCLB) in 2003 and the test scores of the year 2002. Table 11 indicates that this is not the case; while column (D) shows that actual differences in the test scores for the lowest percentiles of the student achievement distribution are large between the years 2002 and 2003, column (E) indicates that these differences almost disappear when NCLB parameters were turned off[46]. Therefore, this table suggests that reallocation of teachers seems to have little effect on the test score performance of low ability students after NCLB.

Finally, it is analyzed how an optimal allocation of teachers for lagging students could have improved their performance. In this regard, schools were divided by the number of classrooms of the same grade, and students were allocated to their best available teacher for a given grade in their school[47]. More specifically, this mechanism allocates first the worst student in a given grade

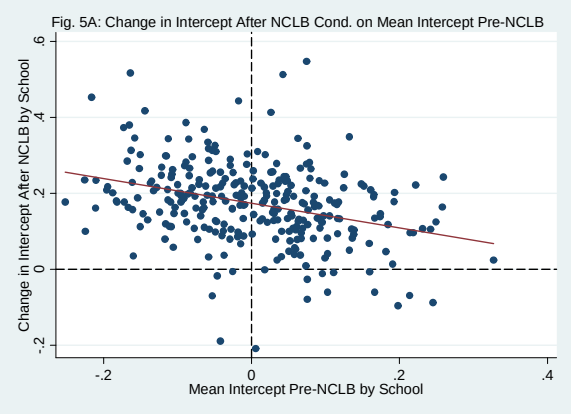
[46] In addition, in terms of teacher turn over, the data indicates that while between the years 2001-2002 the proportion of teachers that changed school was 5.4%, for the period 2002-2003 this number only incresed to 7.4%

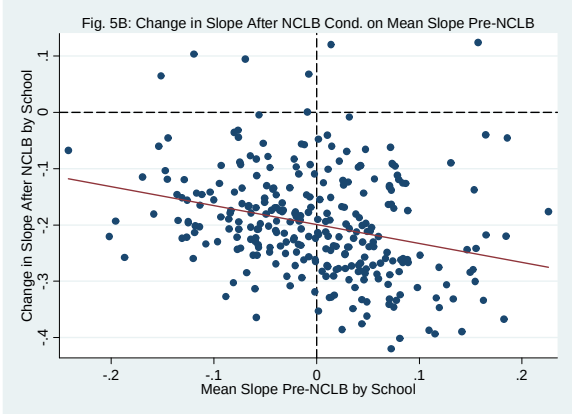
[47] Reallocation of teachers across grades seems not to be substantial after NCLB. For instance, 87% (90%) of the





Figures 5A-B: Change in slopes and intercepts by school conditional on mean test scores pre NCLB.





Figures 6A-B: Change in slopes and intercepts by school conditional on mean intercept and slope pre NCLB.

Evolution of Math Test Scores by Year and Predicted Values without NCLB in Grade 3					
Student Percentile	2002	2003	Pred. 2003 w/o NCLB	Diff. (B)-(A)	Diff. (C)-(A)
	(A)	(B)	(C)	(D)	(E)
10^{th}	-1.24	-0.76	-1.20	0.48	0.04
25^{th}	-0.52	-0.15	-0.58	0.37	-0.06
50^{th}	0.09	0.33	0.09	0.24	0.00
75^{th}	0.81	0.81	0.80	0	-0.01
90^{th}	1.41	1.29	1.40	-0.12	-0.01
Mean	0.11	0.32	0.10	0.21	-0.01
Std. Dev.	0.99	0.81	1.00	0.18	0.01

Table 11: Math test scores for different students percentiles in years 2002 and 2003 in grade 3 and predicted test scores in 2003 after turning off the NCLB parameters.

Best Allocation Mechanism for Low Ability Students				
Difference "Counterfactual" - Actual Performance				
	Student Terciles			
Number of Courses per Grade	T1	T2	T3	All T
2	0.09	0.01	-0.03	0.02
3	0.13	0.02	-0.02	0.04
4 or more	0.15	0.04	0.02	0.07

Table 12: Change in students performance under the best allocation mechanism for low ability students. Differences between counterfactual performance and actual performance were opened by student terciles and number of classrooms per grade in a given school.

and school with the best available teacher (in their school and grade), then the same procedure is applied to the second worst student and so on. The number of students assign to a given teacher cannot exceed the size of the course that originally appeared in the data; when the course is completed that teacher is no longer available, and the teacher - student allocation continues but only considering the remaining available teachers. Table 12 shows the difference between the counterfactual performance (under the described mechanism) and the actual performance, open by student (ability) terciles and the number of classrooms of the same grade in a given school. Results indicate that important increases in lag students performance can be achieved if this allocation mechanism is implemented; however, these counterfactual improvements in students performance are still smaller than the changes that occurred after the implementation of NCLB.

To conclude, model estimation results indicate that after the introduction of NCLB, the mechanism that has driven improvements on less proficient students can be characterized by a combination of an increase in teacher homogeneous effects, and a substantial change in the interaction effects benefitting less advantaged students and hurting the most proficient ones.

8 Conclusions

This paper has shown that teacher - student interactions play a key role when analyzing test scores production function, and the mechanisms behind the outcomes of accountability policies. While most works in the literature do not study complementarity effects, this paper has shown,

teachers that were assign to grade 3 in 2001 (2002) continue in that same grade in 2002 (2003).

for instance, that student - teacher interactions can explain more than 30% of total teacher effectiveness for the 75^{th} percentile student. An analysis of NCLB indicates that lagging students have benefitted with the implementation of this accountability program due to the fact that educators increased their intercepts and decreased their slopes. In this regard, it is established that 60% of the improvement in test scores for low performing kids can be explained by a change in interaction effects. On the one hand, teachers have increased the type of effort that improves the performance of all students irrespective of their characteristics after NCLB; but on the other hand they have also reallocated part of their attention from the most advantaged students to those with lower ability. As a consequence, results indicate that the top third of the student achievement distribution experienced no gains in test scores in the period post NCLB program, while the top quintile shows a decrease of -0.09 of a standard deviation. Moreover, the evidence suggests that the change in student performance post NCLB could not be attributed to a reallocation of teachers within school (conditional on grade).

In addition, this paper shows that identification and consistency of teacher intercepts and slopes can be obtained even when the number of observations per student is small. Moreover, an iterative algorithm is implemented making it possible to circumvent the inconvenience of dealing with high dimensional vectors of fixed effects given that the within-group transformation fails to purge interactive fixed effects.

Future research might consider an analysis of teachers sorting process across schools conditional on their slopes. For instance, it can be important to determine whether high (low) slope teachers tend to move to schools where the achievement level of the student body is higher (lower) than, for example, the state mean. It could also be the case that accountability policies may prevent many teachers to work in schools that better fit their characteristics if those schools are frequently sanctioned. In addition, the econometric strategy implemented in this article can be used in future research to study firm-worker interaction effects. For instance, the Longitudinal Employer-Household Dynamics (LEHD) program database integrates information from state unemployment insurance data with Census Bureau economic and demographic data. Such integrated data permits the construction of longitudinal information on workforce composition at the employer level. Therefore, using LEHD data, it can be possible to determine whether the production function of a subgroup of firms relies more on complementarity effects than others (e.g. technology firms vs. fast food firms).

References

[1] Aaronson, Daniel, Lisa Barrow, and William Sander. (2007). "Teachers and Student Achievement in the Chicago Public High Schools", *Journal of Labor Economics*, 25 (1),95-135.

[2] Abowd, J. M. and Kramarz, F.: 1999, "The Analysis of Labor Markets using Matched Employer-Employee Data", in O. C. Ashenfelter and D. Card (eds), Handbook of Labor Economics, Vol. 3B, Elsevier Science B.V., chapter 40, pp. 2629–2710

[3] Ahn, S., Lee, Y., and Schmidt, P, (2001). "GMM Estimation of Linear Panel data Models with Time-Varying Individual Effects," *Journal of Econometrics*, 101, 219-255.

[4] Ahn, T. and Vigdor, J. (2009) "Does No Child Left Behind Have Teeth? Examining the Impact of Federal Accountability Sanctions in North Carolina," Duke University Working Paper.

[5] Ahn, T. and Vigdor, J. (2010). "The Impact of Incentives on Effort: Teacher Bonuses in North Carolina," Prepared for the PEPG Conference Merit Pay: Will It Work? Is It Politically Viable?

[6] Arcidiacono, P., Foster, G., Goodpaster, N., and Kinsler, J. (2011). "Estimating Spillovers using Panel Data, with an Application to the Classroom" Working Paper Duke University

[7] Bai, J. (2009). "Panel Data Model with Interactive Fixed Effects," *Econometrica*, Vol. 77, No 4, 1229-1279

[8] Ballou, D. (2009). "Test Scaling and Value-Added Measurement," *Education Finance and Policy* Vol. 4, No. 4, Pages 351-383

[9] Brandt, L., Siow A. and Wang, H. (2009). "Substitution Effects in Parental Investment," IZA, DP No. 4431

[10] Clotfelter, C., Ladd, H. and Vigdor, J. (2006). "Teacher-Student Matching and the Assessment of Teacher Effectiveness" NBER Working Paper 11936.

[11] Condie, S., Lefgren, L. and Sims, D. (2011). "Heterogenous Match Quality and Teacher Value-Added: Theory and Empirics" Working Paper Brigham Young University.

[12] Cooley, J. and Tracczynski, J. (2011). "Spare the Rod? The Effect of No Child Behind on Failing Schools," Working Paper University of Wisconsin, Madison.

[13] Dee, T. (2004). "Teachers, Race, and Student Achievement in a Randomized Experiment," *Review of Economics and Statistics*, 86:1 (2004), 195-210.

[14] Dee, T. and Jacob, B. (2011). "The Impact of No Child Left Behind on Student Achievement," *Journal of Policy Analysis and Management*, Vol. 30, No. 3, 418-446

[15] Greene, W. (2004). "Fixed Effects and the Incidental Parameters Problem in the Tobit Model," *Econometric Reviews*, 23, (2): 125-148

[16] Hanushek, E. and Raymond, M. (2004). "Does School Accountability Lead to Improved Student Performance," NBER Working Paper No.10591.

[17] Hanushek, E., Kain, J., O'Brien, D. and Rivkin, S. (2005). "The Market for Teacher Quality" NBER Working Paper No. 11154

[18] Hanushek, E. and Rivkin, S. (2010). "The Quality and Distribution of Teachers Under the No Child Left Behind Act," *Journal of Economic Perspectives*, Volume 24, Number 3, pp: 133-150

[19] Harris, D. (2009). "Would Accountability Based on Teacher Value Added be Samrt Policy? An Examination of the Statistical Properties and Policy Alternatives," *Education Finance and Policy,* Vol. 4, No. 4, Pages 319-350

[20] Hendry, D. and Srba, F. (1977). "The Properties of Autoregressive Instrumental Variables Estimators in Dynamic Systems," *Econometrica*, Vol. 45, pp. 969-990.

[21] Holmstrom, B. and Milgrom, P. (1991). "Multi-task Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law Economics & Organization*. Vol 7 (0). P24-52. Special Issue, 1991.

[22] Ishii, J. and Rivkin, S. (2009). "Impediments to the Estimation of Teacher Value Added," *Education Finance and Policy* Vol. 4, No. 4, Pages 520-536

[23] Jacob, B., Lefgren, L. and Sims, D. (2009). "The Persistence of Teacher-Induced Learning Gains," Journal of Human Resources, 2010, 45(4): 915-943.

[24] Kane, T., and Staiger, D. (2002). "The Promise and Pitfalls of Using Imprecise School Accountability Measures," The Journal of Economic Perspectives, Vol. 16, No.4, pp 91-114.

[25] Kinsler, J. (2011). "Beyond Levels and Growth: Estimating Teacher Value-Added and its Persistence," Working Paper University of Rochester

[26] Koedel, C. and Betts, J. (2007). "Re-Examining the Role of Teacher Quality in the Educational Production Function" Working Paper University of Missouri.

[27] Koretz, D. (2002). "Limitations in the Use of Achievement Test Measures of Educators' Productivity," *Journal of Human Resources*, 37(4), pp: 752-777.

[28] Ladd, H. and Lauen, D. (2010). "Status versus Growth: The Distributional Effects of School Accountability Policies," *Journal of Policy Analysis and Management*, Vol. 29, No. 3, 426-450

[29] Lee, Y., and Schmidt, P. (1993). "A production frontier model with flexible temporal variation in technical efficiency", in H. K. Fried, K. Lovell and S. Schmidt, eds, *'The Measurement of Productive Efficiency'*, Oxford University Press, New York.

[30] Lee, Y. (2006). "A stochastic production frontier model with group-specific temporal variation in technical efficiency", *European Journal of Operational Research*, 1616-1630

[31] Lockwood, J and McCaffrey, D. (2009). "Exploring Student-Teacher Interactions in Longitudinal Data" *Education Finance and Policy,* Vol. 4, No. 4, Pages 439-467

[32] McCaffrey, D., Sass, T., Lockwood, J. and Mihaly, K. (2009). "The Intertemporal Variability of Teacher Effect Estimates" *Education Finance and Policy*, Vol. 4, No. 4, Fall 2009, p. 572-606.

[33] Macartney, H. (2011). "The Dynamic Effects of Educational Accountability," Working Paper Duke University

[34] Murnane, R. and Papay J. (2010). "Teachers' Views on No Child Left Behind: Support for the Principles, Concerns about the Practices", *Journal of Economic Perspectives*, Volume 24, Number 3, Pages 151–166

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

[36] Neal, D. (2011). "The Design of Performance Pay in Education," NBER Working Paper No. 16710

[37] Neyman, J., and Scott, E. (1948). "Consistent Estimates based on Partially Consistent Observations," *Econometrica* 16, 805-840.

[38] Nye, Barbara, Spyros Konstantopoulos and Larry V. Hedges. (2004). "How large are teacher effects?" *Educational Evaluation and Policy Analysis,* 26:237-257.

[39] Reardon, S. and Raudenbush, S. (2009). "Assumptions of Value-Added Models for Estimating school effects," *Educational Evaluation and Policy Analysis*, Vol. 4, No. 4, pp 492-519.

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

[41] Reback, R.,Rockoff, J., and Schwartz, H. (2011). "Under Pressure: Job Security, Resource Allocation, and Productivity Under NCLB" NBER, Working Paper No. 16745

[42] Rivkin, S., Hanushek, E. and Kain, J. (2005)"Teachers, Schools, and Academic Achievement," *Econometrica*, 73 (2), 417-458

[43] Rockoff, J. (2004). "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, pp. 247-252.

[44] Rothstein, J. (2009). "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables," *Education Finance and Policy,* Vol. 4, No. 4, Pages 537-571

[45] Rothstein, J. (2010). "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quaterly Journal of Economics*, 125 (1)

[46] Sargan, J. (1964). "Wages and Prices in the United Kingdom: A Study in Econometrics Methodology" *in Econometric Analysis of National Economic Planning*, ed. by P. G. Hart, G. Mill, and J. K. Whitaker. London, Butterworths, 25-54.

[47] Vigdor, J. (2008). "Teacher Salary Bonuses in North Carolina," Working Paper Duke University

[48] Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge, Massachussets, London, England.

9 Appendix A

9.1 Proof of Identification and Consistency

For ease of exposition consider 1 student that is observed only two periods. In each period, he/she has different teachers (i.e. $j = 1, 2$). Therefore, the non linear least squares problem can be written as follows:

$$\sum_{j=1}^{J=2}(T_{1j} - q_{1j} - q_{2j}S_1 - S_1)^2 \tag{25}$$

Next, it is necessary to normalize the intercept (q_{1j}) and slope (q_{2j}) of one teacher. Therefore, q_{12} and q_{22} are set to be equal A and B. Then, the NLLS problem can be written as follows:

$$(T_{11} - q_{11} - q_{21}S_1 - S_1)^2 + (T_{12} - A - BS_1 - S_1)^2 \tag{26}$$

First the lemmas are listed with the proofs to follow.

Lemma 1: *Concentrating out the student unobserved ability (i.e. S_1) of the original least squares problem results in an optimization problem over the $q's$, that takes the following form:*

$$\min_{q_{11},q_{22}} \frac{[(T_{11} - q_{11})(1 + B) - (T_{12} - A)(1 + q_{21})]^2}{(1 + q_{21})^2 + (1 + B)^2} \tag{27}$$

The nonlinear squares problem now only depends of the $q's$ and the data. Therefore, it is possible to investigate the properties of the estimator of $\mathbf{q}^0 = \{q_{11}^0, q_{21}^0\}$. For ease of notation, define $z(w, \mathbf{q})$ as:

$$z(w, \mathbf{q}) = \frac{[(T_{11} - q_{11})(1 + B) - (T_{12} - A)(1 + q_{21})]^2}{(1 + q_{21})^2 + (1 + B)^2} \tag{28}$$

where $w \equiv \mathbf{T}$.

Lemma 2

$$E\left[\mathbf{z}(\mathbf{w}, \mathbf{q}^0)\right] < E\left[\mathbf{z}(\mathbf{w}, \mathbf{q})\right], \forall\ \mathbf{q}\ \in \Theta, \mathbf{q}_0 \neq \mathbf{q} \tag{29}$$

where Θ is the parameter space.

Theorem 12.1 of Wooldridge (2002) establishes that sufficient conditions for consistency are identification and uniform convergence. Having already established identification, Lemma 3 shows uniform convergence.

Lemma 3

$$\max_{\mathbf{q}\in\Theta}\left|\frac{1}{N}\sum \mathbf{z}(w_i,\mathbf{q})-\mathbf{E}\left[\mathbf{z}(w,\mathbf{q})\right]\right|\xrightarrow{P}0 \tag{30}$$

Consistency then follows from Theorem 12.1 of Wooldridge: $\widehat{\mathbf{q}}\xrightarrow{P}\mathbf{q}_0$

Proof of Lemma 1 The NLLS problem is the following:

$$(T_{11}-q_{11}-q_{21}S_1-S_1)^2+(T_{12}-A-BS_1-S_1)^2 \tag{31}$$

Take first order conditions w.r.t. S_1

$$S_1=\frac{(T_{11}-q_{11})(1+q_{21})+(T_{12}-A)(1+B)}{(1+q_{21})^2+(1+B)^2} \tag{32}$$

If equation (32) is replaced in equation (31), after several steps of algebra it can be obtained:

$$\begin{aligned}\mathbf{z}(\mathbf{w},\mathbf{q}) &= (T_{11}-q_{11})^2+(T_{12}-A)^2\\ &\quad -\left[\frac{(T_{11}-q_{11})(1+q_{21})+(T_{12}-A)(1+B)}{(1+q_{21})^2+(1+B)^2}\right]^2\\ \mathbf{z}(\mathbf{w},\mathbf{q}) &= \frac{\left[(T_{11}-q_{11})(1+B)-(T_{12}-A)(1+q_{21})\right]^2}{(1+q_{21})^2+(1+B)^2}\end{aligned}$$

Therefore, expression (27) is recovered.

QED.

Proof of Lemma 2 Proving identification requires that:

$$E\left[\mathbf{z}(\mathbf{w},\mathbf{q}^0)\right]<E\left[\mathbf{z}(\mathbf{w},\mathbf{q})\right],\forall\ \mathbf{q}\ \in\Theta,\mathbf{q}_0\neq\mathbf{q} \tag{33}$$

where Θ is the parameter space.

Proof:

$$\mathbf{z}(\mathbf{w},\mathbf{q})=\frac{\left[(T_1-q_{11})(1+B)-(T_2-A)(1+q_{21})\right]^2}{(1+q_{21})^2+(1+B)^2} \tag{34}$$

After replacing $\mathbf{T}$ by the data generating process in the previous expression, it can be obtained

(after several steps of algebra):

$$\mathbf{z}(\mathbf{w},\mathbf{q}) = \frac{\left[(1+B)((q_{11}^0-q_{11})+S_1^0(q_{21}^0-q_{21}))\right]^2}{(1+q_{21})^2+(1+B)^2} \tag{35}$$

$$+\frac{\left[\varepsilon_1(1+B)+\varepsilon_2(1+q_{21})\right]^2}{(1+q_{21})^2+(1+B)^2} \tag{36}$$

$$+\frac{2\left[(1+B)((q_{11}^0-q_{11})+S_1^0(q_{21}^0-q_{21}))\right]\left[\varepsilon_1(1+B)+\varepsilon_2(1+q_{21})\right]}{(1+q_{21})^2+(1+B)^2} \tag{37}$$

If expectation is applied to the previous expression, and $E(\varepsilon_j)=0$, $E(\varepsilon_1\varepsilon_2)=0$, $E(\varepsilon_j^2)=\sigma^2$; then when $q_{11}=q_{11}^0$ and $q_{21}=q_{21}^0$, equation (34) is minimized in the truth. Until now, it has been shown identification when considering a pair of teachers:

$$\arg\min_{q_j} \mathbf{Q}_{ij}(q_i^0,q_j) = q_j^0$$

However, it is easy to show that this result can be used to prove identification in the following case (i.e. multiple pair of teachers):

$$\arg\min_{\mathbf{q}} \textstyle\sum_{j=1}^{P}\sum_{i>j}^{P} \mathbf{Q}_{ij}(q_i,q_j)$$

where P denotes the total number of teachers.

QED.

Proof of Lemma 3 To prove consistency is not sufficient to show pointwise convergence in probability (i.e. it is not enough to simply invoke the usual weak law of large numbers at each $\mathbf{q} \in \Theta$). Instead, it is necessary to prove uniform convergence in probability[48]:

$$\max_{\mathbf{q}\in\Theta}\left|\frac{1}{N}\sum \mathbf{z}(w_i,\mathbf{q}) - \mathbf{E}\left[\mathbf{z}(w,\mathbf{q})\right]\right| \xrightarrow{P} 0$$

Theorem 12.1 in Wooldridge states four conditions that the data and $\mathbf{z}$ must satisfy in order for the above condition to hold:

1) Θ is compact

2) For each $\mathbf{q}\in\Theta$, $\mathbf{z}(\cdot,\mathbf{q})$ is Borel measurable on W[49]. Since $\mathbf{z}(\cdot,\mathbf{q})$ is a continuous function of w, it is also Borel measurable.

[48] It is important to mention that consistency requires that the number of students that share a pair of teachers goes to infinity as $N\to\infty$.

[49] Let W be a random vector taking values in $W\subset R^M$ where W denotes the subset of R^M representing the possible values of w.

3) For each $w \in W$, $\mathbf{z}(w, \mathbf{q})$ is continuous on Θ

4) $|\mathbf{z}(w, \mathbf{q})| < b(w)$ for all $\mathbf{q} \in \Theta$, where b is a nonnegative function on W such that $E[b(w)] < \infty$

Points 1 and 2 are satisfied by assumption. Point 3 is straight forward to verify; however, point 4 requires to be proven. Note that $\mathbf{z}(\mathbf{w}, \mathbf{q})$ is always positive so the absolute value can be ignored:

$$\mathbf{z}(\mathbf{w}, \mathbf{q}) = \frac{\left[(T_1 - q_{11})(1 + B) - (T_2 - A)(1 + q_{21})\right]^2}{(1 + q_{21})^2 + (1 + B)^2} \tag{38}$$

Given that $\mathbf{q}$ belongs to a compact space, then it has lower and upper bound values. Therefore, $\mathbf{q}$ can be replaced in (38) by these lower or upper bound values such that the expression is maximize. In addition, if A and B are normalized to be equal 0 (without loss of generality), then it can be obtained (after assuming that $\mathbf{T}$ has a second moment):

$$\begin{aligned}
\mathbf{z}(\mathbf{w}, \mathbf{q}) &= \frac{\left[(T_1 - q_{11}) - T_2(1 + q_{21})\right]^2}{(1 + q_{21})^2 + 1} \\
&< \left[(T_1 - q_{11}) - T_2(1 + q_{21})\right]^2 \\
E(\mathbf{z}(\mathbf{w}, \mathbf{q})) &< \max\left\{E((T_1 - C_{lb})^2, (T_1 - C_{ub})^2)\right\} \\
&\quad + \max\left\{E(T_2^2(1 + D_{lb})^2), E(T_2^2(1 + D_{ub})^2)\right\}
\end{aligned}$$

where C_{lb}, C_{ub}, D_{lb} and D_{ub} denote lower and upper bound values of the $q\prime s$. Then, it can be obtained a function such that $E[b(w)] < \infty$.

QED.

10 Appendix B

The algorithm described in Section 4 needs to be slightly modified when trying to estimate the model expressions that include $\overline{S_g}$ (i.e. mean classroom student fixed effect). The reason is that it is no longer possible to solve for S_i in equation (15). In this regard, step 1 remains similar; however, step 2 is changed and an additional step is included in order to update $\overline{S_g}$. Then, the algorithm becomes:

- **Step 1':** Using the initial guesses of the student fixed effects, calculate $Y_{ijg}^{(m)} = T_{ijg} - S_i^{(m)}$ and solve the least squares problem:

$$\{q_{1j}^{'(m)}, a_{2j}^{'(m)}, \lambda^{'(m)}\} = \arg \min_{e_{1j}, e_{2j}} \sum_{i=1}^{N} \sum_{j=1}^{J} \left(Y_{ijg}^{(m)} - q_{ij} - S_i^{(m)} \, a_{2j} - \lambda \overline{S}_g^{(m)} \right)^2 \tag{39}$$

- **Step 2':** Using $q_{ij}^{'(m)}$, $a_{2j}^{'(m)}$ and $\lambda^{'(m)}$, calculate $S_i^{(m+1)}$ using the following expression (where $j \in i$ denotes all the teachers of student i):

$$\frac{\sum_{j \in i} \left(Test_{ijg} - q_{ij}^{'(m)} - \lambda^{'(m)} \overline{S}_g^{(m)} \right)}{\sum_{j \in i} 1 + a_{2j}^{'(m)}} = S_i^{(m+1)} \tag{40}$$

Now, an additional step is required in order to update $\overline{S}_g^{(m)}$

- **Step 3':** Calculate $\overline{S}_g^{(m+1)}$ using $S_i^{(m+1)}$

- **Step 4'**: Repeat steps 1, 2 and 3 until convergence of the parameters.

To sum up, the main differences between this algorithm and the previous one rely on adding step 3'; and the fact that in step 2' instead of using the first order condition of the least squares problem with respect to S_i to concentrate out this variable, it is used use directly equation (13), the implicit assumption in this case is that the mean of the error term for a given student is 0, $\overline{\varepsilon_{ig}} = 0$. Finally, it is important to emphasize that if step 2 is replaced by step 2' in the algorithm presented in Section 4, results are almost the same as it is shown in the Monte Carlo experiments in Appendix C.

11 Appendix C

Monte Carlo evidence obtained from estimating the two model versions (i.e. including $\overline{T}_{-1g}$ or $\overline{S_g}$) using the algorithm described in appendix B. See subsection 4.3 for a description on how the data was created. Results indicate that the econometric strategy performs quite well for both model equations.

Monte Carlo Evidence							
Model 1: $Test_{ijg} = q_{1j} + a_{2j}S_i + \lambda\overline{T}_{-1g}S_i + S_i + \varepsilon_{ijg}$							
	Teacher Intercepts			Teacher Slopes			
True Parameters	$q_{11}= -1$	$q_{12}= -0.5$	$q_{13}= 1.5$	$a_{21}= -0.5$	$a_{22}= 0.15$	$a_{23}= 0.35$	$\lambda = 0.5$
Estimates	-1.002	-0.498	1.499	-0.488	0.146	0.342	0.490
($\sigma_\varepsilon= 0.25$)	(0.027)	(0.029)	(0.035)	(0.027)	(0.030)	(0.033)	(0.053)
Model 2: $Test_{ijg} = q_{1j} + a_{2j}S_i + \lambda\overline{S_g}S_i + S_i + \varepsilon_{ijg}$							
Estimates	-0.987	-0.499	1.487	-0.496	0.142	0.354	0.478
($\sigma_\varepsilon= 0.25$)	(0.059)	(0.036)	(0.068)	(0.050)	(0.033)	(0.068)	(0.134)

Table C1: Monte Carlo Evidence when estimating two different versions of students test scores production function using the algorithm described in Appendix B. The reported parameters were obtained after estimating each model 100 times, and then taking the average. In parenthesis, it is presented the standard deviation of the coefficients.

12 Appendix D

Evolution of Average Math Test Scores by Year: Grade 3					
	2000	2001	2002	2003	2004
	Full Sample: All North Carolina				
Mean	250.5	250.6	251.4	253.3	253.3
Observations	101429	101969	100652	102042	101598
	Reduced Sample: 5 Biggest Counties				
Math	250.7	251.2	252.1	253.9	253.9
Observations	28710	28726	28670	29334	29543

Table D1: Evolution of math test scores for grade 3 in selected years. Test scores range from 218 - 276 (scores are supposed to be comparable across years). Full sample includes all counties in North Carolina, while reduced sample only includes the 5 biggest counties in the state (i.e. Cumberland, Wake, Forsyth, Guildford and Charlotte-Mecklenburg).

13 Appendix E

Estimation Results				
	Math		Reading	
	Intercept (q_{1j})	Slope (a_{2j})	Intercept (q_{1j})	Slope (a_{2j})
Std. Dev.	0.198	0.148	0.140	0.111
λ	0.023		0.0001	
	(0.005)		(0.005)	
$\beta * I(\exp r < 1yr)$	-0.055		-0.033	
	(0.007)		(0.008)	
L-R test (P-Value)	0.001		0.001	

Table E1: Standard deviation and correlation of teacher fixed effects (i.e. intercepts $\{q_{1jMath}, q_{1j\,\mathrm{Re}\,ad}\}$ and slopes $\{a_{2jMath}, a_{2j\,\mathrm{Re}\,ad}\}$) recovered from the estimation of equation (17) with the only difference that $\overline{S_g}$ was included in the equation instead of $\overline{T}_{-1g}$. Likelihood ratio (L-R) tests results that analyze the joint significance of teachers slopes are also reported. See equation (17) to interpret the parameters λ and β.

14 Appendix F

	Evolution of Math Test Scores							
	Year							
Student Percentile	2001	2002	2003	2004	2001	2002	2003	2004
	Grade 4				Grade 5			
10^{th}	-1.30	-1.19	-0.74	-0.85	-1.28	-1.18	-0.98	-0.88
25^{th}	-0.74	-0.63	-0.17	-0.29	-0.69	-0.69	-0.39	-0.39
50^{th}	-0.61	0.05	0.39	0.39	-0.10	0	0.29	0.29
75^{th}	0.73	0.84	0.96	0.96	0.68	0.78	0.88	0.87
85^{th}	1.07	1.18	1.30	1.30	1.17	1.17	1.17	1.17
90^{th}	1.30	1.52	1.52	1.52	1.37	1.47	1.47	1.47
95^{th}	1.75	1.86	1.75	1.75	1.76	1.86	1.76	1.76
Mean	0	0.10	0.37	0.37	0	0.10	0.23	0.28
Std. Dev.	1	1	0.88	0.9	1	1.02	0.91	0.91

Table F1: Evolution of math test scores pre (2001-2002) and post (2003-2004) NCLB for different percentiles of students achievement in grades 4 and 5. Standard deviations in each year are also reported. Test scores were normalized with respect to the mean and standard deviation of year 2001.

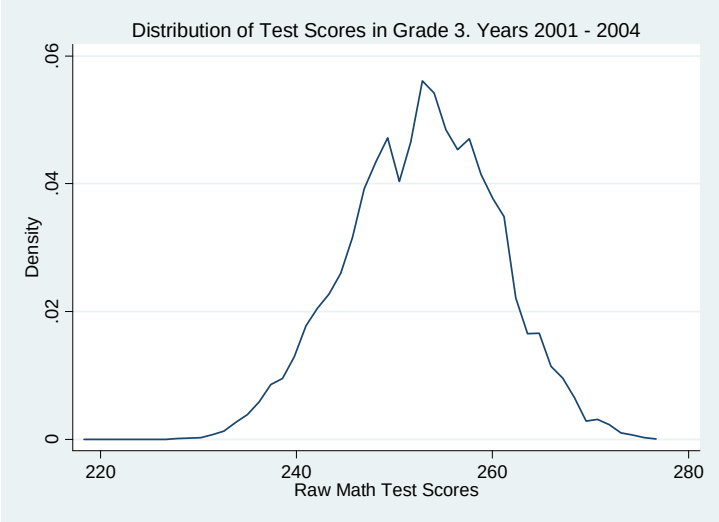


Figure F1: Kernel distribution of raw math test scores in grade 3. Years 2001 -2004

Students Charac. Pre-Post NCLB (Grade 3)			
	2001	2002	2003
Reduced Price Lunch	40.2	41.8	44.3
Parents Education (College)	36.3	36.3	34.6
Minority	46.4	49.6	51.1

Table F2: Students characteristics pre and post NCLB in grade 3: proportion of students with reduced priced lunch, proportion of students with parents with a college degree or more, proportion of minorities.

Teachers Charac. Pre-Post NCLB (Grade 3)			
	2001	2002	2003
Experience less than 1 year	11.6	11.1	13.2
Gender Male	5.2	5	6.9
Race White	79.7	84.7	79.4

Table F3: Teachers characteristics pre and post NCLB in grade 3: proportion of teachers with less than one year of experience, proportion of male teachers, proportion of white teachers.

15 Appendix G

Estimation Results Equation (24)				
ρ	π	φ	δ	β
0.177	0.015	-0.209	0.100	-0.059
(0.003)	(0.003)	(0.002)	(0.001)	(0.006)

Table G1: Estimation results of equation 24 but with the only difference that teachers adjustment post NCLB is forced to be the same for everyone after NCLB (i.e. ρ and φ no longer change by school)

Estimation Results Equation (23)				
ρ	π	φ	δ	β
0.181	0.007	-0.194	0.091	-0.059
(0.003)	(0.003)	(0.002)	(0.001)	(0.006)

Table G2: Estimation results of equation 23 (i.e. $\overline{S_g}$ is used instead of $\overline{T}_{-1g}$) but with the only difference that teachers adjustment post NCLB is forced to be the same for everyone after NCLB (i.e. ρ and φ no longer change by school)