ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-792

Identification of Non-Additive Fixed Effects Models: Is the Return to Teacher Quality Homogeneous?

Jinyong Hahn
UCLA

John D. Singleton
University of Rochester
& NBER

Nese Yildiz
University of Rochester

Panel or grouped data are often used to allow for unobserved individual heterogeneity in econometric models via fixed effects. In this paper, we discuss identification of a panel data model in which the unobserved heterogeneity both enters additively and interacts with treatment variables. We present identification and estimation methods for parameters of interest in this model under both strict and weak exogeneity assumptions. The key identification insight is that other periods' treatment variables are instruments for the unobserved fixed effects. We apply our proposed estimator to matched student-teacher data used to estimate value-added models of teacher quality. We show that the common assumption that the return to unobserved teacher quality is the same for all students is rejected by the data. We also present evidence that No Child Left Behind-era school accountability increased the effectiveness of teacher quality for lower performing students.

VERSION: June 2023

Suggested citation: Hahn, Jinyong, John D. Singleton, and Nese Yildiz. (2023). Identification of Non-Additive Fixed Effects Models: Is the Return to Teacher Quality Homogeneous?. (EdWorkingPaper: 23-792). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/kfh3-1278

Identification of Non-Additive Fixed Effects Models: Is the Return to Teacher Quality Homogeneous?*

Jinyong Hahn
UCLA

John D. Singleton
University of Rochester
& NBER

Neşe Yildiz
University of Rochester

June 2023

Abstract

Panel or grouped data are often used to allow for unobserved individual heterogeneity in econometric models via fixed effects. In this paper, we discuss identification of a panel data model in which the unobserved heterogeneity both enters additively and interacts with treatment variables. We present identification and estimation methods for parameters of interest in this model under both strict and weak exogeneity assumptions. The key identification insight is that other periods' treatment variables are instruments for the unobserved fixed effects. We apply our proposed estimator to matched student-teacher data used to estimate value-added models of teacher quality. We show that the common assumption that the return to unobserved teacher quality is the same for all students is rejected by the data. We also present evidence that No Child Left Behind-era school accountability increased the effectiveness of teacher quality for lower performing students.

JEL Codes: C12, C14, C31, C36, C52, H75, I21, J24
Keywords: *Panel data, fixed effects, teacher quality, school accountability*

*This Version: Thursday 15^{th} June, 2023. Hahn: Department of Economics, UCLA, 8283 Bunche Hall, Mail Stop: 147703, Los Angeles, CA 90095, Email: hahn@econ.ucla.edu. Singleton: Department of Economics, University of Rochester, 230 Harkness Hall, P.O. Box: 270156, Rochester, NY 14627, Email: john.singleton@rochester.edu. Yildiz: Department of Economics, University of Rochester, 222 Harkness Hall, P.O. Box: 270156, Rochester, NY 14627, Email: nese.yildiz@rochester.edu. Nese Yildiz has received financial support from the NSF through grant SES-1918985. Lena Harris provided excellent research assistance.

1 Introduction

Panel or grouped data are often used to allow for unobserved heterogeneity in econometric models via fixed effects. For instance, panel data addresses the possible endogeneity of treatment when selection is based on fixed (e.g. time-invariant) unobservables. In other applications, estimates of fixed effects themselves are of interest. These include estimates of firms' and workers' unobserved productivities using employer-employee matched datasets (e.g. Abowd et al. 1999) and estimates of teachers' unobserved quality using student-teacher matched datasets (e.g. Chetty et al. 2014a).

Classic fixed effects models separate the unobservables into the additive sum of scalar individual heterogeneity α_i – termed fixed effects – and an error term U_{it}. The fixed effects are time (t) invariant and allowed to be correlated with treatment variables X_{it}, while U_{it} is uncorrelated with treatment. As the fixed effects enter these models only in an additively separable way, they are easy to difference out (Chamberlain, 1984; Hsiao, 2014); the "within" transformation establishes identification and provides one estimator.

In this paper, we present conditions for identification of models in which fixed effects enter additively as well as interact with covariates, such as treatment status. As a result, the standard technique of differencing out α_i is no longer valid. The existence of such interactions can have important economic implications: treatment effects will depend on unobserved heterogeneity and the marginal effect of a change in unobserved heterogeneity will vary with treatment. Estimates of treatment effects and of fixed effects will be biased if it is incorrectly presumed that interactions between unobserved heterogeneity and observed variables are null.

To preview our results and their intuition, suppose, as is often assumed in the literature, that the structural equation relating X_{it} and α_i to outcome Y_{it} excludes X_{is} for some $s \neq t$ and that $\mathbb{E}\left[U_{it}|X_{it}, X_{is}\right] = \mathbb{E}\left[U_{it}|X_{it}\right] = 0$.[1] Then, X_{is} can be viewed as an instrument for the unobserved α_i if $\mathbb{E}\left[\alpha_i|X_{it}, X_{is}\right]$ varies with X_{is}. In particular, variation in X_{is} leads to variation in $\mathbb{E}\left[\alpha_i|X_{it}, X_{is}\right]$ holding X_{it} constant. Our key insight is that this variation can be exploited to develop identification and estimation strategies for parameters of interest for a class of non-additive fixed effects models.

[1] This condition holds under both strict and weak exogeneity (pre-determinedness) assumptions.

We present identification results informed by this insight. Our results obtain in "short" panels and they preserve features of additive fixed effects model identification: they do not rely on distributional assumptions about the shape of α_i or place restrictions on the unobserved heterogeneity's relationship with treatment variables. We establish identification results for extensions of our baseline model, such as allowing for more than one treatment variable and for higher order terms of the treatment variable interacted with the fixed effects. We also extend the results to allow for the inclusion of covariates in the model. We then present a non-linear transformation of the model that eliminates α_i. This transformation serves two purposes: First, we derive a conditional moment restriction using the transformation that provides the basis for estimation of non-additive fixed effects models by linear IV regression. Second, as in Holtz-Eakin et al. (1988), this transformation can be used to extend our identification results to the case where the regressors are only pre-determined or weakly exogenous.

As an empirical application, we apply our proposed estimator to matched student-teacher administrative data used to estimate value-added models of teacher quality (e.g. Kane and Staiger 2008; Chetty et al. 2014a). Our data are from the North Carolina Education Research Data Center and we focus on math and reading scores of 4th and 5th grade students. We first show that the common assumption that the return to unobserved teacher quality is the same for all students is rejected by the data. In fact, a counterfactual one standard deviation increase in teacher quality in reading is estimated to be 15% *more* effective for a student one standard deviation below average in their prior score. The estimates also indicate that the return to teacher quality is lower for economically disadvantaged and underrepresented minority students. We show these findings have meaningful implications for estimates of individual teachers' qualities and for how equitably teacher quality is distributed. Ignoring heterogeneity leads to overestimating the quality of teachers assigned to highly disadvantaged classrooms.

We then consider the question of whether interactions between unobserved teacher quality and student characteristics reflect workplace features of education production. To do this, we ask whether and how variation in accountability pressure—incentives linked to student performance on standardized exams—shifts the return to unobserved teacher quality. We do this using No Child Left Behind-era accountability policy in a difference-in-differences framework that leverages the timing of pressure. We show that accountability pressure induced by failure to meet Adequate Yearly Progress targets caused meaningful increases in both subjects in the effectiveness of teacher quality for lower performing students. Combined with the direct effect on teacher quality, the net result of this heterogeneity is that the policy effect of accountability on test scores is over twice as large—about 0.07σ in reading—for a student one standard deviation below average in their prior score.[2]

1.1 Literature and Outline

Our paper contributes to the recent literature on panel data models where parameter heterogeneity is present in both the intercept term and slope coefficients. See e.g. Chamberlain (1992); Robertson and Symons (1992); Pesaran and Smith (1995); Durlauf et al. (2001); Browning and Carro (2007). These models are often called random coefficients models and Arellano and Bonhomme (2012) and Graham and Powell (2012) are recent papers studying identification of models of this type in short panels. The models we consider in this paper can be viewed as random coefficient models with a unique parsimonious structure. Specifically, the intercept and the slope coefficient of the treatment variable are both functions of the same scalar unobserved individual heterogeneity. Our paper is thus closer in spirit to Evdokimov (2010), which considers identification of panel data models that do not assume a particular functional relationship between regressors and fixed unobserved heterogeneity.

The parsimonious structure of our model allows us to entertain a different set of conditions for identification than those considered by Arellano and Bonhomme (2012) and Graham and Powell (2012). Arellano and Bonhomme (2012) focus on identification of the probability distribution of random coefficients, while Graham and Powell (2012) examine estimation of their expectation. Our identification results are based on model primitives, reflecting the more recent trends in the econometric literature. Though the model we analyze is in principle less general than the one Evdokimov (2010) studies in its linearity, our identification results and proposed estimator are intuitive and clarify precisely the types of variation in observables needed to achieve identifica-

[2]This calculation assumes the student is assigned to a teacher one standard deviation better than the average teacher.

tion.[3] Moreover, estimation of the class of models we consider is straightforward and requires only appropriately transforming the data and then running a linear IV (e.g. 2SLS) regression to carry out.

Most of the literature, Arellano and Bonhomme (2012) and Evdokimov (2010) included, focuses on the case of strictly exogeneity and it is sometimes not obvious how an extension can be made to the case where the regressors are only weakly exogenous. By considering transformations based on a few adjacent periods, we are able to use an intuition similar to the one in Arellano and Bond (1991) and accommodate the case in which regressors are only pre-determined. Our paper thus makes a contribution by establishing identification results for non-additive fixed effects models with weakly exogenous as well as strictly exogenous regressors.

Our empirical findings contribute to the large literature on measuring and assessing the importance of teacher quality (e.g. Rivkin et al. 2005; Kane and Staiger 2008; Chetty et al. 2014a,b; Koedel and Rockoff 2015). This body of work relies on administrative datasets grouping students and teachers in classrooms to recover estimates of individual teachers' qualities as their value-added to student test scores. This setup embodies an assumption—the return to teacher quality is the same for all students—that we show is rejected by the data. Our paper is thus related to Ahn et al. (2020), who allow for match effects via random coefficients.[4] Finally, our findings identify a new channel supporting the effect of No Child Left Behind's accountability provisions on student learning: teacher quality becomes more effective for the students targeted by the policy.[5]

The remainder of the paper is structured as follows. In Section 2, we introduce the model and the identification results under the assumption that the regressors are strictly exogeneous. Section 3 discusses how controls can be added to the baseline model. In Section 4, we use a transformation to derive conditional moment restrictions suitable for estimation. Section 5 then

[3] We also establish identification for models that include higher order terms of, multiple, and interactions between treatment variables.

[4] Ahn et al. (2020) specify and estimate a joint distribution of the random coefficients (which represent distinct teacher skills or qualities), whereas we identify match effects assuming a scalar teacher quality that maps into learning in a context- and student-specific way and whose distribution we do not require any assumptions about.

[5] Related work on the effectiveness of accountability, NCLB and otherwise, includes Hanushek and Raymond (2005); Ahn and Vigdor (2014); Deming et al. (2016); Hollinger (2021); Mansfield and Slichter (2021). Figlio and Loeb (2011) summarize the literature on school accountability.

presents our empirical application to matched student-teacher data. Finally, Section 6 presents identification results when the regressors are only weakly exogenous or pre-determined, while Section 7 concludes. Proofs not presented in the main text are collected in the appendix.

2 Model

2.1 Baseline Model

In this section we introduce the baseline model and discuss the parameters of interest. We will discuss extensions later. In this baseline model there is a single covariate X_t (which is the main explanatory variable of interest) in each time period t. It has two effects on the outcome: one effect is the same across individuals; the second effect varies across individuals because X_t interacts with unonserved individual fixed effect. In particular, outcome Y_t is determined by

$$Y_{it} = \tilde{\beta}_{0*} + \tilde{\alpha}_i + X_{it}\tilde{\beta}_{1*} + X_{it}\tilde{\beta}_{2*}\tilde{\alpha}_i + U_{it}, \tag{1}$$

where $\tilde{\alpha}$ is a random variable denoting individual specific unobserved heterogeneity, $(\tilde{\beta}_{0*}, \tilde{\beta}_{1*}, \tilde{\beta}_{2*})$ are non-random parameters and U_{it} represents additional unobservables. We are going to assume the number of periods T is small and fixed. In fact, often we are going to assume $T = 2$. The identification results we provide will be based on the observationally equivalent model given by

$$Y_{it} = \alpha_i + X_{it}\beta_{1*} + X_{it}\beta_{2*}\alpha_i + U_{it}. \tag{2}$$

Before we discuss our identification approach we first illustrate that these two models (1) and (2) are observationally equivalent. We do this in two steps. Let $\tilde{c}_* := \mathbb{E}[\tilde{\alpha}]$, and let $\alpha' := \tilde{\alpha} - \tilde{c}_*$, be the demeaned version of the fixed effect. Then the model given in equation (1) is equivalent to

$$Y_{it} = \beta'_{0*} + \alpha'_i + X_{it}\beta'_{1*} + X_{it}\beta'_{2*}\alpha'_i + U_{it},$$

with $\beta'_{0*} := \tilde{\beta}_{0*} + \tilde{c}_*$, $\beta'_{1*} := \tilde{\beta}_{1*} + \tilde{\beta}_{2*}\tilde{c}_*$, and $\beta'_{2*} := \tilde{\beta}_{2*}$. Furthermore, this model is equivalent to the model given in equation (2) with $\alpha_i := \alpha'_i + \beta'_{0*}$, $\beta_{1*} := \beta'_{1*} - \beta'_{2*}\beta'_{0*}$, and $\beta_{2*} = \beta'_{2*}$.

Below, we provide sufficient conditions for the identification of β_{1*}, β_{2*}, $\mathbb{E}[\alpha_i]$ and $\mathbb{E}[\alpha_i|X_{it} = x]$. These are the parameters that show up in evaluation of important marginal/treatment effects. In particular, consider the model given by equation (1). In that model the potential outcomes that we would observe if X_{it} is exogenously set to x and x', respectively, are

$$Y_{it}(x) = \tilde{\beta}_{0*} + \tilde{\alpha}_i + x\tilde{\beta}_{1*} + x\tilde{\beta}_{2*}\tilde{\alpha}_i + U_{it},$$
$$Y_{it}(x') = \tilde{\beta}_{0*} + \tilde{\alpha}_i + x'\tilde{\beta}_{1*} + x'\tilde{\beta}_{2*}\tilde{\alpha}_i + U_{it},$$

so that the effect on individual i from exogenously changing their X_{it} value from x to x' equals

$$Y_{it}(x') - Y_{it}(x) = (x' - x)\tilde{\beta}_{1*} + (x' - x)\tilde{\beta}_{2*}\tilde{\alpha}_i.$$

For this model the average treatment effect is

$$\mathbb{E}\left[Y_{it}(x') - Y_{it}(x)\right] = (x' - x)\tilde{\beta}_{1*} + (x' - x)\tilde{\beta}_{2*}\mathbb{E}\left[\tilde{\alpha}_i\right].$$

If we consider the model given in equation (2) instead, the potential outcomes and average treatment effect become, respectively

$$Y_{it}(x) = \alpha_i + x\beta_{1*} + x\beta_{2*}\alpha_i + U_{it},$$
$$Y_{it}(x') = \alpha_i + x'\beta_{1*} + x'\beta_{2*}\alpha_i + U_{it},$$
$$\mathbb{E}\left[Y_{it}(x') - Y_{it}(x)\right] = (x' - x)\beta_{1*} + (x' - x)\beta_{2*}\mathbb{E}\left[\alpha_i\right].$$

Since $\alpha_i = \tilde{\alpha}_i + \tilde{\beta}_{0*}$, $\beta_{1*} = \tilde{\beta}_{1*} - \tilde{\beta}_{2*}\tilde{\beta}_{0*}$, and $\beta_{2*} = \tilde{\beta}_{2*}$,

$$(x' - x)\left(\beta_{1*} + \beta_{2*}\mathbb{E}\left[\alpha_i\right]\right) = (x' - x)\left(\tilde{\beta}_{1*} + \tilde{\beta}_{2*}\mathbb{E}\left[\tilde{\alpha}_i\right]\right),$$

therefore, when β_{1*}, β_{2*} and $c_* := E[\alpha]$ are identified, the average treatment effect is identified. In addition, for individuals whose initial treatment value at time t, X_t, equals x, the average ceteris paribus effect of changing their treatment value from x to x' is given by

$$\mathbb{E}\left[Y_{it}(x') - Y_{it}(x)|X_t = x\right] = (x' - x)\beta_{1*} + (x' - x)\beta_{2*}\mathbb{E}\left[\alpha_i|X_t = x\right]. \tag{3}$$

Note that since $\mathbb{E}[\alpha|X_t]$ may depend on t, our framework allows for the average effect of changing the treatment value from x to x' for the subpopulation of individuals whose initial treatment value in period t equals x to depend on the t.

2.1.1 Empirical Application: Estimating Teacher Quality

In this subsection, we preview the empirical application of our baseline model. A large literature estimates unobserved teacher quality as a teacher's value-added to student test scores (e.g. Kane and Staiger 2008; Chetty et al. 2014a). This is based on a panel data model where education production is additively separable into the contribution of student-level inputs, captured by student covariates (including their lagged test score); the value-added or quality of the teacher; and an error term.

A representative setup is given by:

$$y_{ijg} = f(y_{ig-1}) + \gamma x_{ig} + \alpha_j + \epsilon_{ijg} \tag{4}$$

The dependent variable in this equation is the test score of student i, who is taught by teacher j, in grade g. y_{ig-1} is the student's score in the prior grade, while x_{ig} is a vector of other observed student covariates that may include sex, race, ethnicity, and (proxies for) economic advantage. Test scores may be linear in the prior score or $f()$ may be specified as a polynomial, e.g. a cubic. α_j summarizes the quality of teacher j while ϵ_{ijg} represents remaining determinants of learning that are unobserved. Importantly, teacher quality, α_j, is also unobserved to the econometrician and the goal of estimation is to recover reliable estimates of these parameters in addition to estimates of $f()$ and γ.[6]

In this paper, we consider the case where education production is no longer linear in teacher

[6] Empirical Bayes techniques are common to reduce noise in individual teacher effect estimates, which are only consistent as class size grows (see e.g. Koedel and Rockoff 2015; Gilraine et al. 2021. Note that these techniques can be combined with the results and methods in this paper in the same "two step" fashion.

quality because the return is heterogeneous across students. This heterogeneity is summarized by the inclusion of a new term that is the interaction of teacher quality with observed student covariates. For example:

$$y_{ijg} = f(y_{ig-1}) + \gamma x_{ig} + \alpha_j + \beta y_{ig-1}\alpha_j + \upsilon_{ijg} \tag{5}$$

This equation is equivalent to the standard setup above, except for the addition of the $\beta y_{ig-1}\alpha_j$. This term captures the heterogeneity in the return to α_j with respect to students' current learning level, with β governing the magnitude and nature of the heterogeneity. In the case where $\beta > 0$, students with higher prior scores will benefit relatively more from an increase in teacher quality, while students with lower prior scores benefit more when $\beta < 0$ (all else held equal).

In Section 5, we apply the results in this paper to matched student-teacher data to estimate models like equation (5) that relax the assumption that the return to teacher quality is common across students.

2.2 Identification of Baseline Model

For identification, we maintain that $\mathbb{E}[|Y_{it}|] < \infty$, $\mathbb{E}[|X_{it}|] < \infty$, and that $\mathbb{E}[|\alpha_i|] < \infty$. The main assumption we make, in addition to these maintained assumptions, is a form of strict exogeneity assumption.

Assumption 1 *For each $t = 1, 2$, and $s \neq t$ $\mathbb{E}[U_{it}|X_{it}, X_{is}] = \mathbb{E}[U_{it}|X_{it}] = 0$.*

Strict exogeneity assumptions are commonly made for identification of panel data models. While this assumption is restrictive, it may be more believable in the context of fixed (short) T panels considered in this paper.[7] In Section 6, we discuss identification of the baseline model under a pre-determinedness assumption instead.

Our identification approach will rely on first differencing. Specifically, under Assumption 1,

[7] Note that in Section 3, we discuss how to introduce covariates W_{it} into the baseline model. In that version of the model, we require strict exogeneity of X_i conditional on these additional covariates.

for $s \neq t$ we have

$$\mathbb{E}\left[Y_t - Y_s | X_s = x_s, X_t = x_t\right] = \left(x_t - x_s\right)\left(\beta_{1*} + \beta_{2*}\mathbb{E}\left[\alpha | X_s = x_s, X_t = x_t\right]\right). \tag{6}$$

When $x_t \neq x_s$ we can divide both sides of equation (6) by $(x_t - x_s)$ to identify

$$\boxed{\beta_{1*} + \beta_{2*}\mathbb{E}\left[\alpha | X_s = x_s, X_t = x_t\right]}. \tag{7}$$

Note that equation (2) implies that

$$\mathbb{E}\left[Y_t | X_t = x_t, X_s = x_s\right] = \mathbb{E}\left[\alpha | X_s = x_s, X_t = x_t\right] + x_t\beta_{1*} + x_t\beta_{2*}\mathbb{E}\left[\alpha | X_s = x_s, X_t = x_t\right]. \tag{8}$$

Subtracting x_t times the identified object (7) from (8), we identify

$$\boxed{\mathbb{E}\left[\alpha | X_s = x_s, X_t = x_t\right]}. \tag{9}$$

Inspection of (7) leads to our key insight: if $\mathbb{E}\left[\alpha | X_s = x_s, X_t = x_t\right]$ depends on X_{is}, we could use the variation in X_{is} to identify $\boxed{\beta_{1*}}$ and $\boxed{\beta_{2*}}$.

This strategy in essence treats X_{is} with $s \neq t$ as an instrument for the endogenous variable α_i. In order to understand this interpretation, note that we can rewrite (2) as

$$Y_{it} = \mathbb{E}\left[\alpha | X_{is}, X_{it}\right] + X_{it}\beta_{1*} + X_{it}\beta_{2*}\mathbb{E}\left[\alpha | X_{is}, X_{it}\right] + U_{it} + \epsilon_{it,s},$$

where $\epsilon_{it,s} := Y_{it} - \mathbb{E}[Y_{it} | X_{is}, X_{it}]$. Thus, dependence of $\mathbb{E}\left[\alpha | X_s = x_s, X_t = x_t\right]$ on X_{is} plays the role of the relevance condition. On the other hand, the strict exogeneity assumption, together with the fact that X_{is} does not directly enter the structural equation for Y_{it} for $s \neq t$ means that X_{is} is a valid instrument for the endogenous variable α_i. The difference between our method and the standard instrumental variables methods is that α_i is an unobserved variable. The following theorem formalizes this intuition.

Theorem 1 *Suppose* $T = 2$ *and that Assumption 1 holds. Let* $\rho_*(x_1, x_2) = \mathbb{E}\left[\alpha | X_1 = x_1, X_2 = x_2\right]$, *and*

$$\mathcal{A}_1 := \{x_1 : \exists x_2, \tilde{x}_2 \text{ such that } (x_1, x_2), (x_1, \tilde{x}_2) \in \mathrm{Supp}(X_1, X_2) \text{ with } x_1 \neq x_2, x_1 \neq \tilde{x}_2, x_2 \neq \tilde{x}_2, \text{ and } \rho_*(x_1, x_2) \neq \rho_*(x_1, \tilde{x}_2)\},$$

$$\mathcal{A}_2 := \{x_2 : \exists x_1, \tilde{x}_1 \text{ such that } (x_1, x_2), (\tilde{x}_1, x_2) \in \mathrm{Supp}(X_1, X_2) \text{ with } x_1 \neq x_2, \tilde{x}_1 \neq x_2, x_1 \neq \tilde{x}_1, \text{ and } \rho_*(x_1, x_2) \neq \rho_*(\tilde{x}_1, x_2)\}.$$

Suppose that $\mathcal{A}_1 \cup \mathcal{A}_2$ is measurable and has strictly positive probability. Then β_{1*} and β_{2*} are identified.

Proof. Suppose that $\mathbb{P}(\mathcal{A}_2) > 0$, and $a \in \mathcal{A}_2$ with b and c as corresponding to two different values of X_1 as specified in $\mathcal{A}_2$. Then,

$$\frac{\mathbb{E}\left[Y_2 - Y_1 | X_1 = b, X_2 = a\right]}{a - b} = \beta_{1*} + \beta_{2*}\rho_*(b, a), \tag{10}$$

$$\frac{a\mathbb{E}\left[Y_1 | X_1 = b, X_2 = a\right] - b\mathbb{E}\left[Y_2 | X_1 = b, X_2 = a\right]}{a - b} = \rho_*(b, a), \tag{11}$$

$$\frac{\mathbb{E}\left[Y_2 - Y_1 | X_1 = c, X_2 = a\right]}{a - c} = \beta_{1*} + \beta_{2*}\rho_*(c, a), \tag{12}$$

$$\frac{a\mathbb{E}\left[Y_1 | X_1 = c, X_2 = a\right] - c\mathbb{E}\left[Y_2 | X_1 = b, X_2 = a\right]}{a - c} = \rho_*(c, a). \tag{13}$$

Subtracting (12) from (10) yields

$$\frac{\mathbb{E}\left[Y_2 - Y_s | X_s = b, X_2 = a\right]}{a - b} - \frac{\mathbb{E}\left[Y_2 - Y_s | X_s = c, X_2 = a\right]}{a - c} = \beta_{2*}\left(\rho_*(b, a) - \rho_*(c, a)\right).$$

Given equations (11) and (13), we can check if $\rho_*(b, a) - \rho_*(c, a) \neq 0$. Moreover, if $\rho_*(b, a) - \rho_*(c, a) \neq 0$, then β_{2*} is identified. Finally, since $\beta_{1*} + \beta_{2*}\rho_*(b, a)$ is identified, these arguments show that β_{1*} is identified as well. The arguments for the case in which $\mathbb{P}(\mathcal{A}_1) > 0$ are similar. ■

This theorem says that if the support of (X_{i1}, X_{i2}) contains $\{a, b, c\}^2$ (with a, b, c distinct from each other), and if holding X_{i2} (X_{i1}) fixed, varying the value of X_{i1} (X_{i2}) causes a change in $\rho_*(x_1, x_2)$, then we can identify (β_{1*}, β_{2*}). If $T = 2$ and if X_1 and X_2 each take the same two values, however, we cannot use this theorem. The proposition below shows β_{1*} and β_{2*} can still be identified.

Proposition 1 *Suppose $T = 2$ and that Assumption 1 holds. In addition, suppose that $\{a, b\}^2 \subseteq \operatorname{Supp}(X_1, X_2)$ with $a \neq b$. If $\rho_*(b, a) \neq \rho_*(a, b)$, then β_{1*} and β_{2*} are identified.*

Proof. See Appendix. ■

As the proof of Theorem 1 illustrates, $\rho_*(x_1, x_2)$ is identified whenever $x_2 \neq x_1$. Therefore, the assumption that $\rho_*(a, b) \neq \rho_*(b, a)$ is empirically verifiable. The next proposition shows that even if X_{it} takes only (the same) two values in each period, we can still identify β_{1*} and β_{2*} as long as $T \geq 3$.

Proposition 2 *Suppose $T = 3$ and that Assumption 1 holds. In addition, suppose that $\operatorname{Supp}(X_1, X_2, X_3) = \{a, b\}^3$ with $a \neq b$. If $\mathbb{E}[\alpha | X_1 = X_2 = a, X_3 = b] \neq \mathbb{E}[\alpha | X_1 = a, X_2 = X_3 = b]$, then β_{1*} and β_{2*} are identified.*

Proof. See Appendix. ■

Next, we discuss identification of $\mathbb{E}[\alpha | X_t = x]$ once β_{1*} and β_{2*} are identified:

Proposition 3 *Suppose that $\mathbb{E}[U_{it} | X_{it}] = 0$, β_{1*} and β_{2*} are identified, and $\mathbb{P}(X_{it}\beta_{2*} = -1) = 0$. Then $\mathbb{E}[\alpha_i | X_{it}]$ is identified a.s., and $\mathbb{E}[\alpha_i]$ is identified.*

Proof. See Appendix. ■

Finally, if (1) is the preferred way of writing the structural equation, recalling that $\alpha_i = \tilde{\alpha}_i + \tilde{\beta}_{0*}$, we can see that β_{0*} is identified under the location normalization $\mathbb{E}[\tilde{\alpha}_i] = 0$.

The identification approach outlined so far suggests a straightforward estimation strategy. When X_{it} is continuously distributed, however, an alternative estimation procedure based on a moment condition approach might be preferable. In Section 4, we discuss how β_{1*} and β_{2*} can be identified and estimated using this alternative approach.

2.3 Extensions

In the Appendix we discuss some extensions of the model. These include the case of multiple treatment variables and of higher order terms of treatment variables in the structural equation.

Appendix Section A.4 and A.5 consider these cases, respectively. The next section considers the inclusion of covariates. In Section 6, we extend the identification results to the case of weak exogeneity. Finally, we also discuss estimation of the model using unconditional moment restrictions in the Appendix.

3 Covariates

In this section, we discuss how we could add controls to the baseline model. By controls we mean observable variables, denoted W_t, whose partial effects on the outcome are not of interest to the researcher. Typically, controls are used to make exogeneity assumptions more credible.

3.1 Coefficients Depending on Covariates

In this section we consider the following alternative model:

$$Y_t = \tilde{\alpha} + \beta_0(W_t) + X_t\tilde{\beta}_1(W_t) + X_t\beta_2(W_t)\tilde{\alpha} + U_t. \tag{14}$$

For brevity of exposition, we assume $T = 2$. Now suppose for each t and any $s \neq t$, $\mathbb{E}\left[U_t|X_t, W_t, X_s, W_s\right] = \mathbb{E}\left[U_t|X_t, W_t\right] = \mathbb{E}\left[U_t|W_t\right] := \lambda_0(W_t)$. Note that under this assumption W_t is not required to be exogenous. Let $\varepsilon_t := U_t - \mathbb{E}\left[U_t|X_t, W_t, X_s, W_s\right] = U_t - \lambda_0(W_t)$. Substituting this into the outcome equation (14) we get

$$Y_t = \tilde{\alpha} + \beta_0(W_t) + \lambda_0(W_t) + X_t\tilde{\beta}_1(W_t) + X_t\beta_2(W_t)\tilde{\alpha} + \varepsilon_t. \tag{15}$$

Recall that X_t is the treatment variable. The potential outcomes are

$$Y_t(x) = \beta_0(W_t) + \lambda_0(W_t) + \tilde{\alpha} + x\tilde{\beta}_1(W_t) + x\beta_2(W_t)\tilde{\alpha} + \varepsilon_t,$$
$$Y_t(x') = \beta_0(W_t) + \lambda_0(W_t) + \tilde{\alpha} + x'\tilde{\beta}_1(W_t) + x'\beta_2(W_t)\tilde{\alpha} + \varepsilon_t.$$

Individual treatment effect equals

$$Y_t(x) - Y_t(x') = (x - x')\tilde{\beta}_1(W_t) + (x - x')\beta_2(W_t)\tilde{\alpha}.$$

Average treatment effect conditional on W_t equals

$$\mathbb{E}\left[Y_t(x) - Y_t(x') \middle| W_t \right] = (x - x')\tilde{\beta}_1(W_t) + (x - x')\beta_2(W_t)\mathbb{E}\left[\tilde{\alpha}|W_t\right].$$

Now let us define $\alpha := \beta_0(W_t) + \lambda_0(W_t) + \tilde{\alpha}$ and $\beta_1(W_t) := \tilde{\beta}_1(W_t) - \beta_2(W_t)\left[\beta_0(W_t) + \lambda_0(W_t)\right]$. With this notation, the model given in equation (15) can equivalently be written as

$$Y_t = \alpha + X_t\beta_1(W_t) + x_t\beta_2(W_t)\alpha + \varepsilon_t. \tag{16}$$

We now discuss identification of the model (16). To keep the notation simple, we assume that $\text{Supp}(W_t) = \text{Supp}(W_{t-1})$.[8] Also assume that for each w in this common support, $\text{Supp}(X_{t-1}, X_t | W_{t-1} = W_t = w)$ contains the points (a, b) and (a, c) with a, b, c all distinct. Then using the conditional versions of the arguments in the proof of Theorem 1, we can show that for each $w \in \text{Supp}(W_t)$, $\beta_1(w)$ and $\beta_2(w)$ is identified. Next, we write

$$\mathbb{E}\left[Y_t | X_t = x, W_t = w\right] - x\beta_1(w) = \mathbb{E}\left[\alpha | X_t = x, W_t = w\right](1 + x\beta_2(w)).$$

Then if support of (X_t, W_t) is such that $1 + x\beta_2(w) = 0$ occurs with zero probability, $\mathbb{E}\left[\alpha | X_t = x, W_t = w\right]$ will be identified for almost every (x, w). Integrating X_t and W_t out we identify $\mathbb{E}\left[\alpha\right]$.

The last point we make is that, even though we only identified $\beta_1(W_t)$ and $\mathbb{E}\left[\alpha\right]$ as opposed to $\tilde{\beta}_1(W_t)$ and $\mathbb{E}\left[\tilde{\alpha}\right]$, and did not even discuss identification of either $\beta_0(W_t)$ or $\lambda_0(W_t)$, if X_t is the treatment variable, none of this matters for causal analysis. To see this note that potential

[8] Otherwise we can identify the average treatment effect for $w \in \text{Supp}(W_t) \cap \text{Supp}(W_{t-1})$.

outcomes based on equation (16) are equal to

$$
\begin{aligned}
Y_t(x) &= \alpha + x\beta_1(W_t) + x\beta_2(W_t)\alpha + \varepsilon_t, \\
Y_t(x') &= \alpha + x'\beta_1(W_t) + x'\beta_2(W_t)\alpha + \varepsilon_t.
\end{aligned}
$$

Thus, individual treatment effect of changing X_t exogenously from x' to x equals

$$
\begin{aligned}
Y_t(x) - Y_t(x') &= (x - x')\beta_1(W_t) + (x - x')\beta_2(W_t)\alpha \\
&= (x - x')\left(\tilde{\beta}_1(W_t) - \beta_2(w_t)\left(\beta_0(W_t) + \lambda_0(W_t)\right)\right) \\
&\quad + (x - x')\beta_2(W_t)\left(\tilde{\alpha} + \beta_2(w_t)\left(\beta_0(W_t) + \lambda_0(W_t)\right)\right) \\
&= (x - x')\tilde{\beta}_1(W_t) + (x - x')\beta_2(W_t)\tilde{\alpha}.
\end{aligned}
$$

Moreover,

$$
\mathbb{E}[Y_t(x) - Y_t(x')|W_t = w] = (x - x')\beta_1(w) + (x - x')\beta_2(w)\mathbb{E}\left[\alpha|W_t = w\right],
$$

which is identified.

3.2 Linear Controls

Estimation of the model in the previous section will be infeasible if the dimension of W_t is big. For that reason, researchers might prefer controlling for W_t in a linear fashion. To illustrate how this can be done for our model, recall our model as expressed in equation (2)

$$
Y_{it} = \alpha_i + X_{it}\beta_{1*} + X_{it}\beta_{2*}\alpha_i + U_{it}.
$$

Let W_{it} denote the controls. The following assumption is the extension of the usual assumption made by empirical economists when they introduce controls in cross-section settings to short T panel data setting. Before we discuss identification with covariates, we present the strict and weak exogeneity assumptions with covariates:

Assumption 2 ***Strict exogeneity with covariates:*** *For each* t*,* $\mathbb{E}\left[U_{it}|X_{i1}, W_{i1}, ..., X_{iT}, W_{iT}\right] = \mathbb{E}\left[U_{it}|X_{it}, W_{it}\right]$ *a.s.*

Assumption 3 ***Weak exogeneity with covariates:*** *For each* t*,* $\mathbb{E}\left[U_{it}|X_{i1}, W_{i1}, ..., X_{it}, W_{it}\right] = \mathbb{E}\left[U_{it}|X_{it}, W_{it}\right]$ *a.s.*

Assumption 4 *(Covariates)*

(i) $\mathbb{E}\left[\alpha_i|X_{it}, W_{it}\right] = W_{it}^{\top}\kappa_{1*} + \phi_t(X_{it})$ *a.s.;*

(ii) $\mathbb{E}\left[U_{it}|W_{it}, X_{it}\right] = W_{it}^{\top}\kappa_{2*}$ *a.s..*

This assumption allows the conditional expectation of each unobservable given treatment and controls to depend on the controls. It requires, however, that each of these conditional expectation functions to be additively separable in treatment and controls, with the part that depends on the controls being linear. Under this assumption, we can write

$$Y_{it} = \xi_i + X_{it}\beta_{1*} + X_{it}\beta_{2*}\xi + W_{it}^{\top}\delta_{1*} + X_{it}W_{it}^{\top}\delta_{2*} + \epsilon_{it}, \tag{17}$$

where $\epsilon_{it} := U_{it} - \mathbb{E}\left[U_{it}|X_{it}, W_{it}\right]$, $\eta_i := \alpha_i - \mathbb{E}\left[\alpha_i|X_{it}, W_{it}\right]$, $\xi_i := \eta_i + \phi(X_{it})$, $\delta_{1*} := \kappa_{1*} + \kappa_{2*}$ and $\delta_{2*} := \kappa_{1*}\beta_{2*}$. Then

$$\begin{aligned} Y_{it} - \mathbb{E}\left[Y_{it}|X_{it}\right] = &\left(W_{it} - \mathbb{E}[W_{it}|X_{it}]\right)^{\top}\delta_{1*} + X_{it}\left(W_{it} - \mathbb{E}\left[W_{it}|X_{it}\right]\right)^{\top}\delta_{2*} \\ &+ \left(\xi_i - \mathbb{E}\left[\xi_i|X_{it}\right]\right)\left(1 + X_{it}\beta_{2*}\right) + \epsilon_{it}. \end{aligned} \tag{18}$$

Note that $\xi_i - \mathbb{E}\left[\xi_i|X_{it}\right] = \eta_i$. Moreover, $\mathbb{E}\left[\eta_i|X_{it}, W_{it}\right] = \mathbb{E}\left[\epsilon_{it}|X_{it}, W_{it}\right] = 0$. Therefore,

$$\begin{bmatrix} \mathbb{E}\left[\left(W_{it} - \mathbb{E}\left[W_{it}|X_{it}\right]\right)\left(\left(\xi_i - \mathbb{E}\left[\xi_i|X_{it}\right]\right)\left(1 + X_{it}\beta_{2*}\right) + \epsilon_{it}\right)\right] \\ \mathbb{E}\left[X_{it}\left(W_{it} - \mathbb{E}\left[W_{it}|X_{it}\right]\right)\left(\left(\xi_i - \mathbb{E}\left[\xi_i|X_{it}\right]\right)\left(1 + X_{it}\beta_{2*}\right) + \epsilon_{it}\right)\right] \end{bmatrix} = \mathbf{0}.$$

As a result, OLS regression of $Y_{it} - \mathbb{E}\left[Y_{it}|X_{it}\right]$ on $W_{it} - \mathbb{E}\left[W_{it}|X_{it}\right]$ and $X_{it}\left(W_{it} - \mathbb{E}\left[W_{it}|X_{it}\right]\right)$ consistently estimates δ_{1*} and δ_{2*}. Since Y_{it}, W_{it} and X_{it} are observed, $\mathbb{E}\left[Y_{it}|X_{it}\right]$ as well as $\mathbb{E}\left[W_{it}|X_{it}\right]$ can be estimated, so that running such an OLS regression is feasible. Once δ_{1*} and

δ_{2*} are identified/estimated, we can subtract $W_{it}^{\intercal}\delta_{1*}$ and $X_{it}W_{it}^{\intercal}\delta_{2*}$ from Y_{it} and we will be back to the scalar covariate model. That is, all of our identification methods can be applied to the model

$$\tilde{Y}_{it} = \xi_i + X_{it}\beta_{1*} + X_{it}\beta_{2*}\xi_i + \epsilon_{it},$$

where $\tilde{Y}_{it} = Y_{it} - W_{it}^{\top}\delta_{1*} - X_{it}W_{it}\delta_{2*}$.

Remark 1 *Note that when X_{it} is binary taking values 0 and 1 only, $\mathbb{E}[Y_{it}|X_{it}]$, $\mathbb{E}[W_{it}|X_{it}]$ and $\mathbb{E}[X_{it}W_{it}|X_{it}]$ are all linear functions and can respectively be estimated by OLS regression of Y_{it} and W_{it} on constant and X_{it}. In fact, by applying Frisch-Waugh-Lovell Theorem OLS regression of $Y_t = (Y_{1t}, Y_{2t}, ..., Y_{nt})^{\top}$ on constant, X_t, W_t and $(XW)_t$, where $(XW)_t = (X_{1t}W_{1t}, X_{2t}W_{2t}, ..., X_{nt}W_{nt})^{\top}$ will yield consistent estimators of δ_1 and δ_2.*

4 Estimation

In this section, we derive conditional moment restrictions that provide the basis for estimation via linear IV regression. For ease of exposition, we first adopt the framework where Xs are strictly exogenous, X_{it} contains a single regressor or treatment variable, and $T = 2$. However, the transformation applied is easily generalized beyond these cases, as appropriate. We also use this transformation later when extending the results to the case of weak exogeneity in Section 6.

Given $Y_{it} = \alpha_i + X_{it}\beta_{1*} + X_{it}\beta_{2*}\alpha_i + U_{it}$, it follows that:

$$\frac{Y_{it} - X_{it}\beta_{1*}}{1 + X_{it}\beta_{2*}} = \alpha_i + \frac{U_{it}}{1 + X_{it}\beta_{2*}} \tag{19}$$

Differencing this expression for $t = 1$ from $t = 2$, we further obtain

$$\frac{Y_{i2} - X_{i2}\beta_{1*}}{1 + X_{i2}\beta_{2*}} - \frac{Y_{i1} - X_{i1}\beta_{1*}}{1 + X_{i1}\beta_{2*}} = \frac{U_{i2}}{1 + X_{i2}\beta_{2*}} - \frac{U_{i1}}{1 + X_{i1}\beta_{2*}}. \tag{20}$$

Strict exogeneity then implies that

$$\mathbb{E}\left[\left.\frac{Y_{i2} - X_{i2}\beta_{1*}}{1 + X_{i2}\beta_{2*}} - \frac{Y_{i1} - X_{i1}\beta_{1*}}{1 + X_{i1}\beta_{2*}}\right| X_{i1}, X_{i2}\right] = 0, \tag{21}$$

which further implies that

$$\mathbb{E}\left[\left(1 + X_{i1}\beta_{2*}\right)\left(Y_{i2} - X_{i2}\beta_{1*}\right) - \left(1 + X_{i2}\beta_{2*}\right)\left(Y_{i1} - X_{i1}\beta_{1*}\right)\right| X_{i1}, X_{i2}] = 0. \tag{22}$$

This equation can be simplified to

$$\mathbb{E}\left[\left(Y_{i2} - Y_{i1}\right) - \left(X_{i2} - X_{i1}\right)\beta_{1*} - \left(X_{i2}Y_{i1} - X_{i1}Y_{i2}\right)\beta_{2*}\right| X_{i1}, X_{i2}] = 0.$$

Therefore, one may hope to use the conditional expectation

$$\mathbb{E}\left[\left(Y_{i2} - Y_{i1}\right) - \left(X_{i2} - X_{i1}\right)\beta_{1} - \left(X_{i2}Y_{i1} - X_{i1}Y_{i2}\right)\beta_{2}\right| X_{i1}, X_{i2}] = 0 \tag{23}$$

as a basis for estimation. Note that this equation is linear in parameters β_{1*} and β_{2*} and α_i does not appear following the initial differencing. $Y_2 - Y_1$ is the dependent variable and the equation contains two right hand side variables: the differences in Xs and $X_2Y_1 - X_1Y_2$. This latter term is endogenous. Following the identification argument earlier and reflected in this moment restriction, identification is obtained from the interaction between X_1 and X_2, an instrument which is "excluded" from the structural equation. The conditional moment restriction in equation (23) thus provides the basis for estimation of β_{1*} and β_{2*} via linear IV regression. $\mathbb{E}\left[\alpha_i | X_{it}\right]$ and $\mathbb{E}\left[\alpha_i\right]$ can then be estimated using equation (19) above.

In the empirical application that follows, we use this conditional moment restriction to estimate a model that features multiple regressors that interact with the unobserved heterogeneity (including one that is continuously distributed) and also includes other covariates that do not interact with the unobserved heterogeneity.

5 Empirical Application: Heterogeneity in the Return to Teacher Quality

In this section, we apply our results to examine and test for heterogeneity in the return to teacher quality across students. To do so, we adapt our proposed estimator to the classroom setting and marshal student-teacher matched administrative from North Carolina. We focus our attention on end-of-grade math and reading scores in 4th and 5th grades in school years 2002-3 to 2008-9. The Data Appendix describes the data and sample construction in detail; Table A1 presents summary statistics.

We begin by estimating the following equation for each subject pooling both grades and all years:

$$y_{ijt} = \gamma x_{it} + \alpha_{jt} + \beta \alpha_{jt} z_{it} + \upsilon_{ijt} \tag{24}$$

Note that this equation is similar to equation (5) previewed earlier, but differs in a couple respects. First, we drop the grade notation and add time (t) subscripts to reflect pooling data across school years. Second, we interact teacher quality with a vector of student covariates, z_{it}. We examine results when only student i's lagged score interacts with teacher quality as well as models that additionally include indicators for economic disadvantage and underrepresented minority (i.e. Black or Hispanic) status in z_{it}. x_{it} then includes the elements in z_{it} as well as lagged score squared and cubed (recall $f()$ in equation (5)), indicators for female, Asian or other non-White race/ethnicity, and whether flagged as an English learner, special education, or gifted. If the commonplace assumption that the return to teacher quality is the same for all students is true, then we expect to fail to reject that $\beta = 0$.

To estimate equation (24), we first residualize y_{ijt} with respect to the covariates not in z_{it} (i.e. those that do not interact with teacher quality). Denote by $\dot{y}_{ijt}$ the residualized test score. We then apply the transformation from the previous section with respect to the classroom average, $\bar{\dot{y}}_{jt}$.[9] In the case where z_{it} contains one variable (e.g. only lagged score), we obtain the following

[9] In principle, however, the transformation could be applied to all student pairs in each classroom.

conditional moment restriction:

$$E[(\dot{y}_{ijt} - \bar{\dot{y}}_{jt}) - \gamma(z_{it} - \bar{z}_{jt}) - \beta(\bar{\dot{y}}_{jt} z_{it} - \dot{y}_{ijt} \bar{z}_{jt})|x_{it}, \bar{x}_{jt}] = 0 \tag{25}$$

where $\bar{z}_{jt}$ collects classroom average characteristics among teacher j's students. This expectation forms the basis of the estimator we use—2SLS. $\bar{\dot{y}}_{jt} z_{it} - \dot{y}_{ijt} \bar{z}_{jt}$ is an endogenous regressor. The instruments we use for our main results are the other elements of x_{it} not in z_{it}, e.g. whether female; class averages of the covariates $\bar{x}_{jt}$; and student-class average interactions, e.g. $x_{it}\bar{x}_{jt}$. For models that include economic disadvantage and underrepresented minority status in z_{it}, the estimating equation includes several additional terms.[10] We cluster standard errors by teacher.

Table 1 presents results for both math and reading. The results allowing for heterogeneity in the return to teacher quality are presented side-by-side with results from estimating models that assume homogeneity. The results show that assumption of common returns to teacher quality is rejected by the data. Column (2) reports estimates for math and shows that the interaction on lagged score and teacher quality is -0.05 and statistically significant. The economic interpretation of this is that a one standard deviation increase in teacher quality would be 5% more effective for a student who is one standard deviation below average in their prior score. Column (3) adds additional teacher quality interactions to the model. The estimates show that an increase in teacher quality is about 3% less effective for an economically disadvantaged student and 6% less effective for an underrepresented minority student in math. Columns (4) through (6) report parallel estimates for reading and show that the economic significance of the interactions with teacher quality is even larger. A one standard deviation improvement in teacher quality would be 16% more effective for a student one standard deviation below average, but 12% less effective for a minority student.

These results in Table 1 show that not all students in a classroom benefit to the same degree by improvements in teacher quality and that, all else equal, lower performing, advantaged, and non-minority students benefit relatively more. These results have important implications for estimates of how teacher quality is distributed and of individual teachers' qualities. On the first

[10] We write out the full estimating equation for this case in the Appendix.

Table 1: Estimates of Education Production Function

	Math score			Reading score		
	(1)	(2)	(3)	(4)	(5)	(6)
Lagged score	0.82***	0.83***	0.82***	0.79***	0.79***	0.79***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Lagged score × Teacher quality		-0.05***	-0.07***		-0.15***	-0.16***
		(0.00)	(0.00)		(0.01)	(0.01)
Econ. disadv.	-0.07***	-0.07***	-0.07***	-0.09***	-0.09***	-0.08***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Econ. disadv. × Teacher quality			-0.03***			-0.04***
			(0.01)			(0.01)
URM	-0.06***	-0.06***	-0.06***	-0.09***	-0.09***	-0.09***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
URM × Teacher quality			-0.06***			-0.12***
			(0.01)			(0.01)
Female	0.00	0.00	0.00	0.03***	0.03***	0.03***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Asian	0.11***	0.10***	0.10***	0.01**	0.01**	0.01**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
other race/ethnicity	-0.03***	-0.03***	-0.03***	-0.04***	-0.04***	-0.03***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Teacher quality μ	0.03	0.03	0.03	0.04	0.04	0.04
Teacher quality σ	0.25	0.25	0.26	0.19	0.20	0.21

N =544,546 student-years (13,747 unique teachers). Standard errors clustered by teacher. All models also control for the square and cubic of lagged score, and indicators of limited English proficiency, special education, and gifted status.

point, a standard deviation increase in the classroom share of economic disadvantage is associated with a 0.003 increase in how much reading teacher quality is overestimated by the homogeneous model. For many teachers, the bias can be economically large: Appendix Table A3 shows that, when heterogeneity is ignored, reading teacher quality is underestimated by more than 0.02 for 21% of the sample and overestimated by more than 0.02 for 22% of the sample.[11]

5.1 Does Accountability Pressure Affect the Return to Teacher Quality?

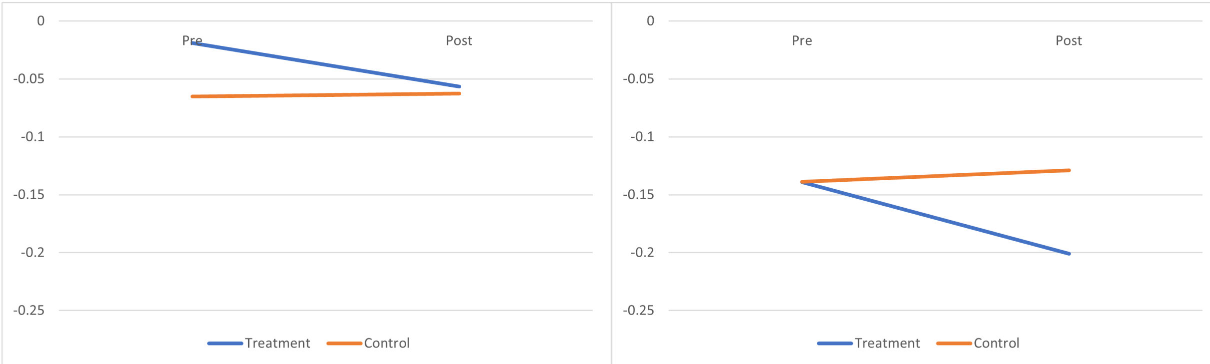
The previous results show that the data reject the assumption that the return to teacher quality is the same for all students. In this section, we examine whether and how the return varies with working conditions. In particular, we examine whether test-based accountability pressure under No Child Left Behind (NCLB) shifts the return and whether the shift is in the direction implied by the policy's goals. The specific aspect of NCLB we focus on is failure to make Adequate Yearly Progress (AYP).

We examine accountability-induced shifts in returns in a difference-in-differences framework. The basic idea is to estimate equation (5) on those schools belonging to each of the four cells—pre-treatment, pre-control, post-treatment, and post-control—and then difference. To implement this idea, we create subsamples for each "cohort" τ from 2004 to 2008.[12] We code as "treated" those schools that missed AYP for the first time the prior year and so experience accountability pressure in school year τ. We code as "control" those schools who have not yet failed AYP—but who eventually will.[13] This variation is summarized in Table A4.

[11]In reading and math, teachers whose quality is overestimated tend to serve more disadvantaged and minority classrooms, while teachers who quality is underestimated tend to serve higher prior achieving classrooms. See the cross tabulations in Appendix Tables A2 and A3 for details.

[12]Note that 2004 is the first year a school could be treated, while schools treated in 2009 have no valid control schools that were first treated in 2010.

[13]Schools that previously failed AYP are excluded from subsequent subsamples and thus do not re-enter the sample as later controls.



(a) Math (b) Reading

Figure 1: Impact of Accountability Pressure on Lagged score × Teacher quality

The difference-in-differences estimating equation is given by:[14]

$$
\begin{aligned}
\dot{y}_{ijt} - \bar{\dot{y}}_{jt} =& \gamma_0(z_{it} - \bar{z}_{jt}) + \beta_0(\bar{\dot{y}}_{jt} z_{it} - \dot{y}_{ijt}\bar{z}_{jt}) + ... + \epsilon^{\tau}_{ijt} \\
&+ \mathbf{1}[t \geq \tau][\gamma_1(z_{it} - \bar{z}_{jt}) + \beta_1(\bar{\dot{y}}_{jt} z_{it} - \dot{y}_{ijt}\bar{z}_{jt})] \\
&+ Treat^{\tau}_j[\gamma_2(z_{it} - \bar{z}_{jt}) + \beta_2(\bar{\dot{y}}_{jt} z_{it} - \dot{y}_{ijt}\bar{z}_{jt})] \\
&+ Treat^{\tau}_j \mathbf{1}[t \geq \tau][\gamma_3(z_{it} - \bar{z}_{jt}) + \beta_3(\bar{\dot{y}}_{jt} z_{it} - \dot{y}_{ijt}\bar{z}_{jt})] \qquad (26)
\end{aligned}
$$

where $Treat^{\tau}_j$ is indicator for whether teacher j was treated in year τ.[15] The subsamples are pooled in estimation and weighted by their inverse number of observations.[16] β_0 summarizes the pre-control return to teacher quality, while β_2 is the difference in the return among treatment group schools—before they are treated. β_3 thus represents the main estimates of interest: the change in the return to teacher quality associated with the onset of accountability pressure.

Figure 1 plots estimates from estimating equation (26) for math and reading, respectively. The figure focuses on how the lagged score interactions with teacher quality are impacted by accountability pressure. In math, the figure shows that control schools' interaction is around -0.06 and this level changes only slightly from pre- to post-. However, treatment schools' interaction

[14] Note that we do not write out all of the terms in the equation for when z_{it} contains multiple variables. In our application, z_{it} contains three variables and the estimating equation includes twelve additional terms in total.

[15] We residualize y_{ijt} separately by treatment-control/pre-post cell for each τ in a first step.

[16] Note that observations corresponding to a school-year may appear up to twice in the final pooled dataset, e.g. year 2004 observations for a school that experiences accountability pressure in 2005 will appear in subsample $\tau = 2004$ (as post-control) and in subsample $\tau = 2005$ (as pre-treatment). The weighting is to guard against negative weights arising from heterogeneity in the effects across τ.

is near zero prior to accountability pressure and then drops to about -0.05. The difference-in-differences estimate is reported in Table 2 and is -0.09 and is statistically different from zero. For reading, the figure is similar: control schools' interaction actually increases from pre- to post-, while the onset of accountability likewise drives the interaction to be much more negative for treated schools. The net impact in reading is likewise an effect of -0.07. In words, these results show that in both math and reading accountability pressure causes the relative return to teacher quality to increase for lower ability students. Table (2) reports estimates on the other quadruple interactions in the model, which show no statistically significant effects on the return for economically disadvantaged or underrepresented minority students.

Table 2: Impact of Accountability Pressure on Return to Teacher quality

	Math	Reading
Lagged score × Teacher quality × Treat × Post	-0.09***	-0.07*
	(0.03)	(0.04)
Econ. disadv. × Teacher quality × Treat × Post	0.02	0.02
	(0.06)	(0.07)
URM × Teacher quality × Treat × Post	-0.00	-0.07
	(0.07)	(0.08)

N =176,496 student-years (4,206 unique teachers). Standard errors clustered by school.

The results above provide novel evidence that NCLB-era accountability raised the relative return to teacher quality for lower ability students. What do our results imply for the net impact of accountability pressure? Conceptually, accountability can raise teacher quality—perhaps through an effort channel—in addition to changing how quality maps into students' learning. To examine this, we take the teacher quality estimates implied by the results above and regress them on accountability pressure in a difference-in-differences fashion.[17] These results, reported in Appendix Table A5, show that teacher quality increases by around 0.03 in math and 0.04 in reading.[18] We then consider the net effects of accountability on two students, both assigned to a high-quality teacher (i.e. a teacher one standard deviation better than the average teacher): one who is one standard deviation below average in lagged score and one who is one standard

[17]Specifically, we estimate $\hat{\alpha}_{jt} = \delta Post_t \times Treat_j + \theta_j + \psi_t + e_{jt}$.

[18]Notably, this direct effect of accountability pressure on reading teacher quality is not present when heterogeneity in the return to teacher quality is ignored. See Appendix Table A5.

deviation above average. For the above average student, the net impact of accountability is zero in math and about 0.02σ in reading. This is because the direct effect channel on teacher quality is offset (fully, in the case of math) by teacher quality becoming less effective for them. In contrast, the below average student benefits from both better teacher quality and that teacher quality being more effective: the implied policy impact on the below average student is 0.06σ in math and 0.07σ in reading.

6 Identification Under Weak Exogeneity

In this section we discuss identification of parameters without assuming strict exogeneity. Thus, we replace Assumption 1 with the following assumption:

Assumption 5 *For each $t = 1, 2, ..., T$, the U_{it} in the model* (1) *satisfies* $\mathbb{E}\left[U_{it}|X_i^t\right] = \mathbb{E}\left[U_{it}|X_{it}\right] = 0$*, where* $X_i^t := (X_{i1}, X_{i2}, ..., X_{it})^\top$.

In contrast to the strict exogeneity assumption, this assumption only requires U_{it} to be conditionally mean independent of X_i^{t-1} when we also condition on X_{it}. Thus, U_{it} can arbitrarily depend on future values of treatment even after we condition on the current value of the treatment. Before we start discussing identification under this assumption, we maintain the following additional assumption throughout this section:

Assumption 6 *For each $t = 1, 2, ..., T$,* $\mathbb{E}\left[\left|\frac{U_t}{1+X_2\beta_{2*}}\right|\right] < \infty$.

6.1 Conditional Moment Restrictions

Under Assumption 5, the identification arguments leading up to Theorem 1 as well as those in sections 2.3, A.4, and A.5 no longer work. Similarly, equation (22) no longer holds. We can, however, still use a similar idea underlying equation (21), i.e., the conditional moment condition

$$\mathbb{E}\left[\frac{Y_{it} - X_{it}\beta_{1*}}{1 + X_{it}\beta_{2*}} - \frac{Y_{it-1} - X_{it-1}\beta_{1*}}{1 + X_{it-1}\beta_{2*}}\,\middle|\, X_i^{t-1}\right] = 0$$

to identify the parameters. With $T = 2$, identification of β_* requires that the (β_1, β_2) that solves

$$0 = \mathbb{E}\left[\frac{(X_{i1} - X_{i2})\,\alpha_i}{(1 + X_{i2}\beta_2)(1 + X_{i1}\beta_2)} (\beta_{2*} - \beta_2) + \frac{X_{i1} - X_{i2}}{(1 + X_{i2}\beta_2)(1 + X_{i1}\beta_2)} (\beta_{1*} - \beta_1) \middle| X_{i1} \right] \tag{27}$$

is equal to (β_{1*}, β_{2*}). Because of the nonlinearity, it is difficult to come up with a primitive condition for identification, although we can discuss it in some special cases. For example, if (X_{i1}, X_{i2}) has only four support points $(a,a), (a,b), (b,a), (b,b)$, each associated with positive probability, and if $(1 + a\beta_{2*})(1 + b\beta_{2*}) \neq 0$, it is straightforward to show that requirement (27) becomes equivalent to $(\beta_1, \beta_2) = (\beta_{1*}, \beta_{2*})$ be the only solution to

$$\begin{aligned} 0 &= \frac{(a-b)\,\rho_*(a,b)}{(1+b\beta_2)(1+a\beta_2)} (\beta_{2*} - \beta_2) + \frac{a-b}{(1+b\beta_2)(1+a\beta_2)} (\beta_{1*} - \beta_1), \\ 0 &= \frac{(b-a)\,\rho_*(b,a)}{(1+a\beta_2)(1+b\beta_2)} (\beta_{2*} - \beta_2) + \frac{b-a}{(1+a\beta_2)(1+b\beta_2)} (\beta_{1*} - \beta_1). \end{aligned}$$

After some simplification we can show that identification of β_* is equivalent to

$$0 \neq \det \begin{bmatrix} \rho_*(a,b) & 1 \\ -\rho_*(b,a) & -1 \end{bmatrix} = \rho_*(b,a) - \rho_*(a,b).$$

When $T = 3$, we can exploit

$$0 = \mathbb{E}\left[\frac{Y_{i3} - X_{i3}\beta_{1*}}{1 + X_{i3}\beta_{2*}} - \frac{Y_{i2} - X_{i2}\beta_{1*}}{1 + X_{i2}\beta_{2*}} \middle| X_{i1}, X_{i2} \right]$$

as well. Thus, point identification of β_* means that any (β_1, β_2) that solves

$$0 = \mathbb{E}\left[\frac{(X_{i2} - X_{i3})\,\alpha_i}{(1 + X_{i3}\beta_2)(1 + X_{i2}\beta_2)} (\beta_{2*} - \beta_2) + \frac{X_{i2} - X_{i3}}{(1 + X_{i3}\beta_2)(1 + X_{i2}\beta_2)} (\beta_{1*} - \beta_1) \middle| X_{i1}, X_{i2} \right]$$

is equal to (β_{1*}, β_{2*}). Suppose that $(X_{i1}, X_{i2}, X_{i3}) = \{a, b\}^3$ with each point in this support associated with positive probability. Consider β such that $(1 + a\beta_2)(1 + b\beta_2) \neq 0$, and let $\theta_*(x_1, x_2, x_3) := \mathbb{E}[\alpha | X_1 = x_2, X_2 = x_2, X_3 = x_3]$, a.s. Note that for the case $(X_{i1}, X_{i2}, X_{i3}) =$

(a, b, a), the above equality is equivalent to

$$0 = \frac{(b-a)\,\theta_*\,(a,b,a)}{(1+a\beta_2)\,(1+b\beta_2)}\,(\beta_{2*}-\beta_2) + \frac{b-a}{(1+a\beta_2)\,(1+b\beta_2)}\,(\beta_{1*}-\beta_1)\,.$$

When $(X_{i1}, X_{i2}, X_{i3}) = (b, b, a)$, the above equality is equivalent to

$$0 = \frac{(b-a)\,\theta_*\,(b,b,a)}{(1+b\beta_2)\,(1+a\beta_2)}\,(\beta_{2*}-\beta_2) + \frac{b-a}{(1+X_{i3}\beta_2)\,(1+a\beta_2)}\,(\beta_{1*}-\beta_1)\,.$$

These two equations, after multiplication by $(1+b\beta_2)\,(1+a\beta_2)\,/\,(b-a)$, become

$$0 = \theta_*\,(a,b,a)\,(\beta_{2*}-\beta_2) + (\beta_{1*}-\beta_1)\,,$$

$$0 = \theta_*\,(b,b,a)\,(\beta_{2*}-\beta_2) + (\beta_{1*}-\beta_1)\,.$$

The solution exists as a unique value at (β_{1*}, β_{2*}) if

$$0 \neq \det\begin{bmatrix} \theta_*\,(a,b,a) & 1 \\ \theta_*\,(b,b,a) & 1 \end{bmatrix} = \theta_*\,(a,b,a) - \theta_*\,(b,b,a)\,.$$

Thus, as long as θ_* varies with the value of X_1 we obtain point identification of β_* in this case even if θ_* is symmetric in (x_2, x_3).[19]

Before we end this section, let us consider the $T = 2$ case again. Suppose there exists $E \subseteq \mathrm{Supp}(X_1)$ with $\mathbb{P}(E) > 0$, such that for each $x_1 \in E$, the support of $X_2|X_1 = x_1$ contains at least three distinct points and that E contains at least two distinct values x_1 and $\tilde{x}_1$. Let $f_{X_2|X_1}(x_2|x_1)$ denote the conditional density of X_2 given $X_1 = x_1$. Then evaluating equation (27) at $X_1 = x_1$ and $X_1 = \tilde{x}_1$ for $\tilde{x}_1 \neq x_1$ and $x_1, \tilde{x}_1 \in E$ yields[20]

$$\begin{bmatrix} \int \frac{x_2-x_1}{(1+x_2\beta_2)(1+x_1\beta_2)} f_{X_2|X_1}(x_2|x_1)\mathrm{d}x_2 & \int \frac{(x_2-x_1)\rho_*(x_1,x_2)}{(1+x_2\beta_2)(1+x_1\beta_2)} f_{X_2|X_1}(x_2|x_1)\mathrm{d}x_2 \\ \int \frac{x_2-\tilde{x}_1}{(1+x_2\beta_2)(1+\tilde{x}_1\beta_2)} f_{X_2|X_1}(x_2|\tilde{x}_1)\mathrm{d}x_2 & \int \frac{(x_2-\tilde{x}_1)\rho_*(\tilde{x}_1,x_2)}{(1+x_2\beta_2)(1+\tilde{x}_1\beta_2)} f_{X_2|X_1}(x_2|\tilde{x}_1)\mathrm{d}x_2 \end{bmatrix}$$

[19] If θ_* does not vary with X_1, but varies with X_2, we could use $0 = \mathbb{E}\left[\frac{Y_{i3}-X_{i3}\beta_{1*}}{1+X_{i3}\beta_{2*}} - \frac{Y_{i1}-X_{i1}\beta_{1*}}{1+X_{i1}\beta_{2*}}\,\middle|\, X_{i1}, X_{i2}\right]$ to obtain point identification of β_* in a similar fashion.

[20] In writing these equations we assumed that for each $x_1 \in E$ (27) is well-defined.

If $f_{X_2|X_1}(x_2|x_1) = f_{X_2}(x_2)$ for each $x_1 \in E$ and ρ_* does not depend on X_1 (so that $\rho_*(x_1, x_2) = \rho_*(\tilde{x}_1, x_2)$ for almost all x_2), the second column is a multiple of the first, and β_* is not identified. On the other hand, even if $f_{X_2|X_1}(x_2|x_1)$ and $\rho_*(x_1, x_2)$ both vary with $x_1 \in E$, we cannot rule out that the possibility that the determinant of the above matrix will be 0, although we would expect the set of β values for which the above matrix has zero determinant to be countable. Thus, proper analysis of estimation and inference of the set of β_* satisfying these conditional moment restrictions will have to use tools from the partial identification literature. We leave this for future research.

6.2 Special Case: Binary Treatment

In this section, we discuss identification of parameters under pre-determinedness when X_t is binary taking values a and b, with $b \neq a$, for each t, even when T is small. For this purpose, consider the $T = 2$ case first.

$$\begin{aligned}
\mathbb{E}\left[Y_1|X_1 = a\right] &= \mathbb{E}\left[Y_1|X_1 = X_2 = a\right]\mathbb{P}(X_2 = a|X_1 = a) \\
&\quad + \mathbb{E}\left[Y_1|X_1 = a, X_2 = b)\mathbb{P}(X_2 = b|X_1 = a\right] \\
&= \{\rho_*(a, a) + a[\beta_{1*} + \beta_{2*}\rho_*(a, a)]\}\mathbb{P}(X_2 = a|X_1 = a) \\
&\quad + \{\rho_*(a, b) + a[\beta_{1*} + \beta_{2*}\rho_*(a, b)]\}\mathbb{P}(X_2 = b|X_1 = a) \\
&= \mathbb{E}\left[Y_2|X_1 = X_2 = a\right]\mathbb{P}(X_2 = a|X_1 = a) \\
&\quad + \{\rho_*(a, b) + a[\beta_{1*} + \beta_{2*}\rho_*(a, b)]\}\mathbb{P}(X_2 = b|X_1 = a).
\end{aligned}$$

Therefore,

$$\rho_*(a, b) + a[\beta_{1*} + \beta_{2*}\rho_*(a, b)] = \frac{\mathbb{E}\left[Y_1|X_1 = a\right] - \mathbb{E}\left[Y_2|X_1 = X_2 = a\right]\mathbb{P}(X_2 = a|X_1 = a)}{\mathbb{P}(X_2 = b|X_1 = a)}. \quad (28)$$

Moreover,

$$\mathbb{E}\left[Y_2|X_1 = a, X_2 = b\right] = \rho_*(a, b) + b\left(\beta_{1*} + \beta_{2*}\rho_*(a, b)\right). \quad (29)$$

Combining (28) and (29), we can see that $\boxed{\beta_{1*} + \beta_{2*}\rho_*(a,b)}$ is identified by

$$\frac{\mathbb{E}\left[Y_2|X_1 = a, X_2 = b\right] - \frac{\mathbb{E}[Y_1|X_1=a] - \mathbb{E}[Y_2|X_1=X_2=a]\mathbb{P}(X_2=a|X_1=a)}{\mathbb{P}(X_2=b|X_1=a)}}{b-a} = \beta_{1*} + \beta_{2*}\rho_*(a,b). \tag{30}$$

Next, by subtracting b times (30) from (29), we identify $\boxed{\rho_*(a,b)}$. Repeating these steps for $X_1 = b$ and $X_2 = a$, we can identify $\boxed{\beta_{1*} + \beta_{2*}\rho_*(b,a)}$ and $\boxed{\rho_*(b,a)}$. Now we can test whether $\rho_*(a,b)$ is equal to $\rho_*(b,a)$. If they are not equal, then we can identify $\boxed{\beta_{2*}}$, and then also $\boxed{\beta_{1*}}$.

Next suppose $T = 3$. Note that

$$\begin{aligned}
\mathbb{E}\left[Y_2|X_1 = X_2 = a\right] &= \mathbb{E}\left[Y_2|X_1 = X_2 = a, X_3 = a\right]\mathbb{P}(X_3 = a|X_1 = X_2 = a) \\
&\quad + \mathbb{E}\left[Y_2|X_1 = X_2 = a, X_3 = b\right]\mathbb{P}(X_2 = b|X_1 = X_2 = a) \\
&= \{\theta_*(a,a,a) + a[\beta_{1*} + \beta_{2*}\theta_*(a,a,a)]\}\mathbb{P}(X_3 = a|X_1 = X_2 = a) \\
&\quad + \{\theta_*(a,a,b) + a[\beta_{1*} + \beta_{2*}\theta_*(a,a,b)]\}\mathbb{P}(X_3 = b|X_1 = X_2 = a) \\
&= \mathbb{E}\left[Y_3|X_1 = X_2 = X_3 = a\right]\mathbb{P}(X_3 = a|X_1 = X_2 = a) \\
&\quad + \{\theta_*(a,a,b) + a[\beta_{1*} + \beta_{2*}\theta_*(a,a,b)]\}\mathbb{P}(X_3 = b|X_1 = X_2 = a).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{\mathbb{E}\left[Y_2|X_1 = X_2 = a\right] - \mathbb{E}\left[Y_3|X_1 = X_2 = X_3 = a\right]\mathbb{P}(X_3 = a|X_1 = X_2 = a)}{\mathbb{P}(X_2 = b|X_1 = X_2 = a)} \\
&\quad = \theta_*(a,a,b) + a\left(\beta_{1*} + \beta_{2*}\theta_*(a,a,b)\right).
\end{aligned} \tag{31}$$

We also have

$$\mathbb{E}\left[Y_3|X_1 = X_2 = a, X_3 = b\right] = \theta(a,a,b) + b\left(\beta_{1*} + \beta_{2*}\theta_*(a,a,b)\right). \tag{32}$$

From (31) and (32), we identify

$$\boxed{\beta_{1*} + \beta_{2*}\theta_*(a,a,b)}, \quad \text{and} \quad \boxed{\theta_*(a,a,b)}. \tag{33}$$

Repeating the same steps with $\mathbb{E}[Y_2|X_1 = X_2 = b]$ and $\mathbb{E}[Y_3|X_1 = X_2 = b, X_3 = a]$, we can also identify

$$\boxed{\beta_{1*} + \beta_{2*}\theta_*(b, b, a)}, \quad \text{and} \quad \boxed{\theta_*(b, b, a)}. \tag{34}$$

Then provided $\theta_*(a, a, b) \neq \theta_*(b, b, a)$, we can identify $\boxed{\beta_{2*}}$ and $\boxed{\beta_{1*}}$ from the identified objects (33) and (34).

7 Conclusion

In this paper, we introduced a short T panel data model in which the intercept and the coefficient on treatment variable are both functions of a scalar variable which represents unobserved individual heterogeneity. We provided novel identification results as well as intuitive linear IV estimators for parameters of this model. We also provided identification results for extensions of the model, including multiple treatment variables and in which higher order terms of treatment variables and their interactions with the unobserved individual heterogeneity enter into the structural equation (under strict exogeneity). Finally, we provided sufficient conditions for identification and estimation of parameters in our baseline model assuming that the regressors are only weakly endogenous (pre-determined). Our identification results illustrate clearly that the dependence of the conditional expectation of the unobserved individual heterogeneity on other periods' treatment values holding current period's treatment value fixed is essential to obtain identification.

We applied the results to matched student-teacher data to test the assumption, common in the prior literature, that the return to unobserved teacher quality is the same for all students. In this application, identification leverages that interactions between a student's own characteristics and those of their classmates are excluded from the structural equation. We show that the assumption that the return to teacher quality is homogeneous is rejected by the data in both math and reading: teacher quality is less effective on average for disadvantaged and minority students, all else equal, and its effectiveness decreases with a student's prior test score. Further, we show that exogenous changes in incentives due to No Child Left Behind-era school accountability pressure raised the effectiveness of teacher quality for those students lagging behind.

References

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*(2), 251–333.

Ahn, T., E. Aucejo, J. James, et al. (2020). The importance of matching effects for labor productivity: Evidence from teacher-student interactions. Technical report, Working Paper, California Polytechnic State University.

Ahn, T. and J. Vigdor (2014). The impact of no child left behind's accountability sanctions on school performance: Regression discontinuity evidence from north carolina. Technical report, National Bureau of Economic Research.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies 58*(2), 277–297.

Arellano, M. and S. Bonhomme (2012). Identifying distributional characteristics in random coefficients panel data models. *The Review of Economic Studies 79*(3), 987–1020.

Browning, M. and J. Carro (2007). Heterogeneity and microeconometrics modelling. *Econometric Society Monographs 43*, 47.

Chamberlain, G. (1984). Panel data. *Handbook of econometrics 2*, 1247–1318.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica 60*(3), 567–596.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American economic review 104*(9), 2593–2632.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review 104*(9), 2633–2679.

Deming, D. J., S. Cohodes, J. Jennings, and C. Jencks (2016). School accountability, postsecondary attainment, and earnings. *Review of Economics and Statistics 98*(5), 848–862.

Durlauf, S. N., A. Kourtellos, and A. Minkin (2001). The local solow growth model. *European Economic Review 45*(4-6), 928–940.

Evdokimov, K. (2010). Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. *Department of Economics, Princeton University*.

Figlio, D. and S. Loeb (2011). School accountability. *Handbook of the Economics of Education 3*, 383–421.

Gilraine, M., J. Gu, R. McMillan, et al. (2021). A nonparametric method for estimating teacher value-added. Technical report.

Graham, B. S. and J. L. Powell (2012). Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models. *Econometrica 80*(5), 2105–2152.

Hanushek, E. A. and M. E. Raymond (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management 24*(2), 297–327.

Hollinger, J. (2021). School accountability, test scores, and long-run outcomes.

Holtz-Eakin, D., W. Newey, and H. S. Rosen (1988). Estimating vector autoregressions with panel data. *Econometrica 71*(5), 1371–1395.

Hsiao, C. (2014). *Analysis of panel data* (3 ed.). Econometric Society Monographs. Cambridge University Press.

Kane, T. J. and D. O. Staiger (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Koedel, C. and J. E. Rockoff (2015). Value-added modeling: A review. *Economics of Education Review 47*, 180–195.

Mansfield, J. and D. Slichter (2021). The long-run effects of consequential school accountability.

Pesaran, M. H. and R. Smith (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics 68*(1), 79–113.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

Robertson, D. and J. Symons (1992). Some strange properties of panel data estimators. *Journal of Applied Econometrics 7*(2), 175–189.

A Appendix

A.1 Proof of Proposition 1

As discussed in the text below Proposition 1, $\rho_*(b,a)$ and $\rho_*(a,b)$ are identified since $a \neq b$ and $(a,b),(b,a) \in \text{Supp}(X_1,X_2)$. Then

$$\frac{\mathbb{E}\left[Y_2 - Y_1 | X_1 = b, X_2 = a\right]}{a-b} = \beta_{1*} + \beta_{2*}\rho_*(b,a), \tag{35}$$

$$\frac{\mathbb{E}\left[Y_2 - Y_1 | X_1 = a, X_2 = b\right]}{b-a} = \beta_{1*} + \beta_{2*}\rho_*(a,b), \tag{36}$$

so that

$$\frac{\frac{\mathbb{E}[Y_2 - Y_1|X_1=b,X_2=a]}{a-b} - \frac{\mathbb{E}[Y_2-Y_1|X_1=a,X_2=b]}{b-a}}{\rho_*(b,a) - \rho_*(a,b)} = \beta_{2*}.$$

Since every term on the left side of the above expression is either known or identified, β_{2*} is identified. Then β_{1*} can be identified using equation (35) or (36). ■

A.2 Proof of Proposition 2

Let $\theta_*(x_1,x_2,x_3) := \mathbb{E}\left[\alpha | X_{i1} = x_1, X_{i2} = x_2, X_{i3} = x_3\right]$. Then

$$\mathbb{E}\left[Y_{i3}|X_{i1} = X_{i2} = a, X_{i3} = b\right] - \mathbb{E}\left[Y_{i2}|X_{i1} = X_{i2} = a, X_{i3} = b\right] = (b-a)[\beta_1 + \beta_2\theta_*(a,a,b)],$$

From this equation, we see that $\beta_1 + \beta_2\theta_*(a,a,b)$ is identified. Moreover, using the level equation for period 3 (or period 2 or period 1 potentially), we can identify $\theta_*(a,a,b)$. Similarly,

$$\mathbb{E}\left[Y_{i2}|X_{i1} = a, X_{i2} = X_{i3} = b\right] - \mathbb{E}\left[Y_{i2}|X_{i1} = a, X_{i2} = X_{i3} = b\right] = (b-a)[\beta_1 + \beta_2\theta_*(a,b,b)].$$

This equation identifies both $\beta_1 + \beta_2\theta_*(a,b,b)$ and $\theta_*(a,b,b)$ using period 2 outcome equation. Then the following two equations in two unknowns identifies β_1 and β_2 as long as $\theta_*(a,a,b) \neq \theta_*(a,b,b)$.

$$\frac{\mathbb{E}\left[Y_{i3}|X_{i1} = X_{i2} = a, X_{i3} = b\right] - \theta_*(a,a,b)}{b} = \beta_1 + \beta_2\theta_*(a,a,b),$$

$$\frac{\mathbb{E}\left[Y_{i2}|X_{i1} = a, X_{i2} = X_{i3} = b\right] - \theta_*(a,b,b)}{b} = \beta_1 + \beta_2\theta_*(a,b,b). \qquad ■$$

A.3 Proof of Proposition 3

Since β_{1*} and β_{2*} are assumed to have been identified, under the strict exogeneity assumption we have

$$\mathbb{E}\left[\alpha_i | X_{it} = x\right] = \frac{\mathbb{E}\left[Y_{it} - X_{it}\beta_{1*} | X_{it} = x\right]}{1 + x\beta_{2*}}.$$

Since $\mathbb{E}\left[\alpha_i | X_{it} = x\right]$ is identified for almost every value of X_{it}, $\mathbb{E}\left[\alpha_i\right]$ is also identified. ■

A.4 Interactions of Treatment Variables

In this section, we discuss identification and estimation of the following model in which two distinct treatment variables interact with each other: Consider $T = 4$.

$$Y_t = \alpha + X_{1t}\beta_{11*} + X_{1t}\beta_{12*}\alpha + X_{2t}\beta_{21*} + X_{2t}\beta_{22*}\alpha + X_{1t}X_{2t}\delta_{1*} + X_{1t}X_{2t}\delta_{2*}\alpha + U_t.$$

Consider x_j, x_k, x_l, x_m and t_1, t_2, with $x_j \neq x_k$, $x_l \neq x_m$ and $t_1 \neq t_2$, such that $(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)$ is in the support of $(X_{11}, X_{12}, X_{13}, X_{14}, X_{21}, X_{22}, X_{23}, X_{24})$, and let

$$\lambda_*(X_{11}, X_{12}, X_{13}, X_{14}, X_{21}, X_{22}, X_{23}, X_{24}) := \mathbb{E}\left[\alpha | X_{11}, X_{12}, X_{13}, X_{14}, X_{21}, X_{22}, X_{23}, X_{24}\right].$$

Then

$$\begin{aligned}
&\mathbb{E}\left[Y_4 - Y_3 | X_{11} = x_j, X_{12} = x_k, X_{13} = x_l, X_{14} = x_m, X_{21} = X_{22} = t_1, X_{23} = X_{24} = t_2\right] \\
&\quad = (x_m - x_l)\left[\beta_{11*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\beta_{12*} + t_2\{\delta_{1*} + \delta_{2*}\lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\}\right] \\
&\quad = (x_m - x_l)\left[\beta_{11*} + t_2\delta_{1*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\{\beta_{12*} + t_2\delta_{2*}\}\right]. \qquad (37)
\end{aligned}$$

Since $x_l \neq x_m$, we identify

$$\boxed{\beta_{11*} + t_2\delta_{1*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\left(\beta_{12*} + t_2\delta_{2*}\right)} \qquad (38)$$

from (37). Similarly, we have

$$\begin{aligned}
&\mathbb{E}\left[Y_2 - Y_1 | X_{11} = x_j, X_{12} = x_k, X_{13} = x_l, X_{14} = x_m, X_{21} = X_{22} = t_1, X_{23} = X_{24} = t_2\right] \\
&\qquad = (x_k - x_j)\left(\beta_{11*} + t_1\delta_{1*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\left(\beta_{12*} + t_1\delta_{2*}\right)\right). \qquad (39)
\end{aligned}$$

Since $x_j \neq x_k$, we identify

$$\boxed{\beta_{11*} + t_1\delta_{1*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\left(\beta_{12*} + t_1\delta_{2*}\right)} \qquad (40)$$

from (39). Next, note that the difference of (38) and (40)

$$\begin{aligned}
&\beta_{11*} + t_2\delta_{1*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\{\beta_{12*} + t_2\delta_{2*}\} \\
&\qquad - \Bigg(\beta_{11*} + t_1\delta_{1*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\{\beta_{12*} + t_1\delta_{2*}\}\Bigg) \\
&\qquad\qquad = (t_2 - t_1)\Bigg(\delta_{1*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\delta_{2*}\Bigg).
\end{aligned}$$

Since $t_2 \neq t_1$, this means that

$$\boxed{\delta_{1*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\delta_{2*}} \tag{41}$$

is identified. Subtracting t_1 times (41) from (40), we achieve identification of

$$\boxed{\beta_{11*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\beta_{12*}}. \tag{42}$$

On the other hand,

$$\begin{aligned}
&\mathbb{E}\left[Y_3 - Y_2 | X_{11} = x_j, X_{12} = x_k, X_{13} = x_l, X_{14} = x_m, X_{21} = X_{22} = t_1, X_{23} = X_{24} = t_2\right] \\
&\quad = (x_l - x_k)\left[\beta_{11*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\beta_{12*}\right] \\
&\qquad + (x_l t_2 - x_k t_1)\left[\delta_{1*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\delta_{2*}\right] \\
&\qquad + (t_2 - t_1)\left[\beta_{21*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\beta_{22*}\right].
\end{aligned} \tag{43}$$

Because (41) and (42), which appear in the first two terms on the right of (43), are identified, we can see that this equation identifies

$$\boxed{\beta_{21*} + \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\beta_{22*}}. \tag{44}$$

Now, note that

$$\begin{aligned}
&\mathbb{E}\left[Y_1 | X_{11} = x_j, X_{12} = x_k, X_{13} = x_l, X_{14} = x_m, X_{21} = X_{22} = t_1, X_{23} = X_{24} = t_2\right] \\
&= \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2) \\
&\qquad + x_j\left(\beta_{11*} + \beta_{12*}\lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\right) \\
&\qquad + t_1\left(\beta_{21*} + \beta_{22*}\lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\right) \\
&\qquad + x_j t_1\left(\delta_{1*} + \delta_{2*}\lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)\right).
\end{aligned} \tag{45}$$

Because the last three terms on the right of (45) are multiples of (42), (44), and (41), we can see that

$$\boxed{\lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)} \tag{46}$$

is identified. Next, consider changing exactly one of x_j, x_k, x_l, x_m, t_1 or t_2. Say it is the value of X_3 that is changed from x_l to $\tilde{x}_l$, with $\tilde{x}_l$ different from both x_l and x_m.[21] Going through almost

[21] Obviously, changing all or a subset of components of $(x_j, x_k, x_l, x_m, t_1, t_2)$ would also work, provided that the support of $(X_{11}, X_{12}, X_{13}, X_{14}, X_{21}, X_{22}, X_{23}, X_{24})$ is sufficiently rich.

identical steps we can identify

$$\boxed{\delta_{1*} + \lambda_*(x_j, x_k, \tilde{x}_l, x_m, t_1, t_1, t_2, t_2)\delta_{2*}}, \quad \boxed{\beta_{11*} + \lambda_*(x_j, x_k, \tilde{x}_l, x_m, t_1, t_1, t_2, t_2)\beta_{12*}},$$
$$\boxed{\beta_{21*} + \lambda_*(x_j, x_k, \tilde{x}_l, x_m, t_1, t_1, t_2, t_2)\beta_{22*}}, \quad \boxed{\lambda_*(x_j, x_k, \tilde{x}_l, x_m, t_1, t_1, t_2, t_2)}.$$

Then if $\lambda_*(x_j, x_k, \tilde{x}_l, x_m, t_1, t_1, t_2, t_2) \neq \lambda_*(x_j, x_k, x_l, x_m, t_1, t_1, t_2, t_2)$, which is a testable assumption, $\boxed{\delta_{1*}}$, $\boxed{\delta_{2*}}$, $\boxed{\beta_{11*}}$, $\boxed{\beta_{12*}}$, $\boxed{\beta_{21*}}$ and $\boxed{\beta_{22*}}$ are all identified. Once these parameters are identified, $\boxed{\lambda_*(x_1, x_2, x_3, x_4, t_1, t_2, t_3, t_4)}$ is also identified for each $(x_1, x_2, x_3, x_4, t_1, t_2, t_3, t_4)$ in the support of $(X_{11}, X_{12}, X_{13}, X_{14}, X_{21}, X_{22}, X_{23}, X_{24})$.

A.5 Higher Order Terms

Consider

$$Y_t = \alpha + X_t\beta_{11*} + X_t^2\beta_{12*} + X_t\beta_{21*}\alpha + X_t^2\beta_{22*}\alpha + U_t. \tag{47}$$

We assume $T = 2$ for brevity. We also maintain the strict exogeneity assumption (Assumption 1) in this section. To discuss identification of this model we first assume that there exist x_1, x_2, x_3, x_4 all distinct from each other and distinct from 0 such that $(0, x_1), (0, x_2), (0, x_3), (0, x_4) \in \text{Supp}(X_1, X_2)$. With this assumption, $\mathbb{E}[Y_1|X_1 = 0, X_2 = x_j] = \rho_*(0, x_j)$ for $j = 1, 2, 2, 3$. So for each x such that $(0, x) \in \text{Supp}(X_1, X_2)$, $\rho_*(0, x)$ is identified. Then for $x \neq 0$:

$$\mathbb{E}[Y_2|X_1 = 0, X_2 = x] = \rho_*(0, x) + x[\beta_{11*} + \beta_{21*}\rho_*(0, x)] + x^2[\beta_{12*} + \beta_{22*}\rho_*(0, x)] \iff$$
$$\frac{\mathbb{E}(Y_2|X_1 = 0, X_2 = x) - \rho_*(0, x)}{x} = \beta_{11*} + \beta_{21*}\rho_*(0, x) + x[\beta_{12*} + \beta_{22*}\rho_*(0, x)].$$

Then if the matrix

$$\begin{bmatrix} 1 & \rho_*(0, x_1) & x_1 & x_1\rho_*(0, x_1) \\ 1 & \rho_*(0, x_2) & x_2 & x_2\rho_*(0, x_2) \\ 1 & \rho_*(0, x_3) & x_3 & x_3\rho_*(0, x_3) \\ 1 & \rho_*(0, x_4) & x_4 & x_4\rho_*(0, x_4) \end{bmatrix}$$

is invertible, $\beta_* = (\beta_{11*}, \beta_{12*}, \beta_{21*}, \beta_{22*})$ is identified. Note that since $\rho_*(0, x_j)$ is identified, the invertibility of this matrix is verifiable. Similarly, if the support of (X_1, X_2) is sufficiently rich, in the sense that $(0, x_1), (0, x_2), ..., (0, x_J) \in \text{Supp}(X_1, X_2)$ for $J \geq 2K + 1$, then the parameters in the model

$$Y_t = \sum_{k=1}^{K} X_t^k\beta_{1k*} + \alpha\left[1 + \sum_{k=1}^{K} X_t^k\beta_{2k*}\right] + U_t,$$

is identified.

Next, we discuss identification of the model given in equation (47) without requiring that the support of X_1 contains 0. Instead, we are going to assume we have three periods of data ($T = 3$) and that $(x, x_j, x_k), (x, x_j, x_l) \in \text{Supp}(X_1, X_2, X_3)$ with x, x_j, x_k, x_l all different from each other.

Note that

$$
\begin{aligned}
\mathbb{E}\left[Y_1|X_1=x,X_2=x_j,X_3=x_k\right] &= \theta_*(x,x_j,x_k)+x\left[\beta_{11*}+\theta_*(x,x_j,x_k)\beta_{21*}\right]+x^2\left[\beta_{12*}+\theta_*(x,x_j,x_k)\beta_{22*}\right],\\
\mathbb{E}\left[Y_2|X_1=x,X_2=x_j,X_3=x_k\right] &= \theta_*(x,x_j,x_k)+x_j\left[\beta_{11*}+\theta_*(x,x_j,x_k)\beta_{21*}\right]+x_j^2\left[\beta_{12*}+\theta_*(x,x_j,x_k)\beta_{22*}\right],\\
\mathbb{E}\left[Y_3|X_1=x,X_2=x_j,X_3=x_k\right] &= \theta_*(x,x_j,x_k)+x_k\left[\beta_{11*}+\theta_*(x,x_j,x_k)\beta_{21*}\right]+x_k^2\left[\beta_{12*}+\theta_*(x,x_j,x_k)\beta_{22*}\right].
\end{aligned}
$$

Taking differences and rearranging we get

$$
\begin{aligned}
&\frac{\mathbb{E}\left[Y_2|X_1=x,X_2=x_j,X_3=x_k\right]-\mathbb{E}\left[Y_1|X_1=x,X_2=x_j,X_3=x_k\right]}{x_j-x}\\
&\qquad=\beta_{11*}+\theta_*(x,x_j,x_k)\beta_{21*}+(x_j+x)\left(\beta_{12*}+\theta_*(x,x_j,x_k)\beta_{22*}\right),
\end{aligned}
\tag{48}
$$

and

$$
\begin{aligned}
&\frac{\mathbb{E}\left[Y_3|X_1=x,X_2=x_j,X_3=x_k\right]-\mathbb{E}\left[Y_1|X_1=x,X_2=x_j,X_3=x_k\right]}{x_k-x}\\
&\qquad=\beta_{11*}+\theta_*(x,x_j,x_k)\beta_{21*}+(x_k+x)\left(\beta_{12*}+\theta_*(x,x_j,x_k)\beta_{22*}\right).
\end{aligned}
\tag{49}
$$

Differencing once more yields

$$
\frac{\frac{\mathbb{E}[Y_2|X_1=x,X_2=x_j,X_3=x_k]-\mathbb{E}[Y_1|X_1=x,X_2=x_j,X_3=x_k]}{x_j-x}-\frac{\mathbb{E}[Y_3|X_1=x,X_2=x_j,X_3=x_k]-\mathbb{E}[Y_3|X_1=x,X_2=x_j,X_3=x_k]}{x_k-x}}{x_j-x_k}=\beta_{12*}+\theta_*(x,x_j,x_k)\beta_{22*}. \tag{50}
$$

Since everything on the left side of the above equation is identified, the above equation identifies

$$
\boxed{\beta_{12*}+\theta_*(x,x_j,x_k)\beta_{22*}}. \tag{51}
$$

Plugging (51) into (48) or (49) identifies

$$
\boxed{\beta_{11*}+\theta_*(x,x_j,x_k)\beta_{12*}}. \tag{52}
$$

Finally, plugging both $x\left[\beta_{11*}+\theta_*(x,x_j,x_k)\beta_{21*}\right]$ and $x^2\left[\beta_{12*}+\theta_*(x,x_j,x_k)\beta_{22*}\right]$ into the equation $\mathbb{E}[Y_1|X_1=x,X_2=x_j,X_3=x_k]$ identifies

$$
\boxed{\theta_*(x,x_j,x_k)}. \tag{53}
$$

Repeating these arguments with $X_3=x_l$ instead of x_k, we can identify

$$
\boxed{\beta_{12*}+\theta_*(x,x_j,x_l)\beta_{22*}},\quad \boxed{\beta_{11*}+\theta_*(x,x_j,x_l)\beta_{21*}},\quad \boxed{\theta_*(x,x_j,x_l)}. \tag{54}
$$

Then, as long as $\theta_*(x,x_j,x_l)\neq\theta_*(x,x_j,x_k)$,we can identify $\boxed{\beta_{12*}}$, $\boxed{\beta_{22*}}$, $\boxed{\beta_{11*}}$, and $\boxed{\beta_{21*}}$, using the identified objects (51), (52), (53), and (54).

A.6 Unconditional Moment Restrictions

In practice, some unconditional moment restrictions implied by the conditional moment restrictions discussed in the main text are likely to be adopted as a basis of estimation. In this section we discuss how this could be done both under strict and weak exogeneity assumptions.

A.6.1 Under Strict Exogeneity

Under strict exogeneity, one can adopt

$$\mathbb{E}\left[\begin{array}{c} X_1\left\{(Y_2-Y_1)-(X_2-X_1)\,\beta_{1*}-(X_2Y_1-X_1Y_2)\,\beta_{2*}\right\} \\ X_2\left\{(Y_2-Y_1)-(X_2-X_1)\,\beta_{1*}-(X_2Y_1-X_1Y_2)\,\beta_{2*}\right\} \end{array}\right]=0$$

as a basis of estimation. The identifiability of the β is an empirical matter that can be tested. For example, one can rewrite the above equation as

$$\mathbb{E}\left[\begin{array}{c} X_1\,(Y_2-Y_1) \\ X_2\,(Y_2-Y_1) \end{array}\right]=\mathbb{E}\left[\begin{array}{cc} X_1\,(X_2-X_1) & X_1\,(X_2Y_1-X_1Y_2) \\ X_2\,(X_2-X_1) & X_2\,(X_2Y_1-X_1Y_2) \end{array}\right]\left[\begin{array}{c} \beta_{1*} \\ \beta_{2*} \end{array}\right]$$

and we can see that the identifiability is guaranteed if the matrix

$$\mathbb{E}\left[\begin{array}{cc} X_1\,(X_2-X_1) & X_1\,(X_2Y_1-X_1Y_2) \\ X_2\,(X_2-X_1) & X_2\,(X_2Y_1-X_1Y_2) \end{array}\right]$$

is nonsingular, which can be tested from the data.

We can repeat the same idea for $T=3$ case. We have the conditional moment restriction

$$\mathbb{E}\left[(1+X_s\beta_{2*})\,(Y_t-X_t\beta_{1*})-(1+X_t\beta_{2*})\,(Y_s-X_s\beta_{1*})\right|X_1,X_2,X_3]=0,$$

which delivers the estimating equation

$$\mathbb{E}\left[(Y_t-Y_s)-(X_t-X_s)\,\beta_1-(X_tY_s-X_sY_t)\,\beta_2\right|X_1,X_2,X_3]=0. \tag{55}$$

Because $\mathbb{E}\left[Y_t\right|X_1,X_2,X_3]=\theta_*\left(X_1,X_2,X_3\right)+X_t\beta_{1*}+X_t\theta_*\left(X_1,X_2,X_3\right)\beta_{2*}$, where $\theta_*(X_1,X_2,X_3):=(\alpha|X_1,X_2,X_3)$, this amounts to

$$(X_t-X_s)\,(\beta_{1*}-\beta_1)+(X_2-X_1)\,\theta_*\left(X_1,X_2,X_3\right)(\beta_{2*}-\beta_2)=0.$$

In particular, the identification would be achieved if there exist $(X_1,X_2)=(x_1,x_2,x_3)\,,(x_1',x_2',x_3')$

such that the matrix

$$\begin{bmatrix} x_3 - x_2 & (x_3 - x_2)\,\theta_*\,(x_1, x_2, x_3) \\ x_2' - x_1' & (x_2' - x_1')\,\theta_*\,(x_1', x_2', x_3') \end{bmatrix}$$

is nonsingular. Let's try $(x_1, x_2, x_3) = (a, a, b), (x_1', x_2', x_3') = (a, b, b)$, in which case we have

$$\det \begin{bmatrix} b - a & (b - a)\,\theta_*\,(a, a, b) \\ b - a & (b - a)\,\theta_*\,(a, b, b) \end{bmatrix} = (b - a)^2\,(\theta_*\,(a, b, b) - \theta_*\,(a, a, b)) \neq 0$$

under the earlier condition $\theta_*\,(a, b, b) \neq \theta_*\,(a, a, b)$.

We can also derive unconditional moment restrictions, we can derive a unconditional moment restriction from (55). For example, we can use

$$\mathbb{E} \begin{bmatrix} X_1 \{(Y_2 - Y_1) - (X_2 - X_1)\,\beta_{1*} - (X_2 Y_1 - X_1 Y_2)\,\beta_{2*}\} \\ X_2 \{(Y_2 - Y_1) - (X_2 - X_1)\,\beta_{1*} - (X_2 Y_1 - X_1 Y_2)\,\beta_{2*}\} \\ X_3 \{(Y_2 - Y_1) - (X_2 - X_1)\,\beta_{1*} - (X_2 Y_1 - X_1 Y_2)\,\beta_{2*}\} \\ X_1 \{(Y_3 - Y_2) - (X_3 - X_2)\,\beta_{1*} - (X_3 Y_2 - X_2 Y_3)\,\beta_{2*}\} \\ X_2 \{(Y_3 - Y_2) - (X_3 - X_2)\,\beta_{1*} - (X_3 Y_2 - X_2 Y_3)\,\beta_{2*}\} \\ X_3 \{(Y_3 - Y_2) - (X_3 - X_2)\,\beta_{1*} - (X_3 Y_2 - X_2 Y_3)\,\beta_{2*}\} \end{bmatrix} = 0$$

as a basis of GMM estimation. Identifiability of β_* is a testable restriction which amounts to the question whether the rank of the matrix

$$\mathbb{E} \begin{bmatrix} X_1 (X_2 - X_1) & X_1 (X_2 Y_1 - X_1 Y_2) \\ X_2 (X_2 - X_1) & X_2 (X_2 Y_1 - X_1 Y_2) \\ X_3 (X_2 - X_1) & X_3 (X_2 Y_1 - X_1 Y_2) \\ X_1 (X_3 - X_2) & X_1 (X_3 Y_2 - X_2 Y_3) \\ X_2 (X_3 - X_2) & X_2 (X_3 Y_2 - X_2 Y_3) \\ X_3 (X_3 - X_2) & X_3 (X_3 Y_2 - X_2 Y_3) \end{bmatrix}$$

is equal to 2 or not.

A.6.2 Under Weak Exogeneity

When X_t are continuous, using the conditional moment restrictions might be challenging, especially if the sample size is not large. For this reason, we discuss using unconditional moment restrictions. As was the case in the previous section, the key challenge is to come up with intuitive sufficient conditions for point identification of β_*.

With $T=2$, we can exploit the following moment restrictions:

$$\mathbb{E}\left[\begin{array}{c} \frac{Y_{i2}-X_{i2}\beta_{1*}}{1+X_{i2}\beta_{2*}} - \frac{Y_{i1}-X_{i1}\beta_{1*}}{1+X_{i1}\beta_{2*}} \\ X_{i1}\left(\frac{Y_{i2}-X_{i2}\beta_{1*}}{1+X_{i2}\beta_{2*}} - \frac{Y_{i1}-X_{i1}\beta_{1*}}{1+X_{i1}\beta_{2*}}\right) \end{array}\right] = \left(\begin{array}{c} 0 \\ 0 \end{array}\right). \tag{56}$$

Let $\beta := (\beta_1, \beta_2)^\top$. Then whether $\beta_* = (\beta_{1*}, \beta_{2*})^\top$ is identified relative to β by the moment conditions given above depends on whether

$$A_2^{\text{pre}}(\beta) := \mathbb{E}\left[\begin{array}{cc} \frac{(X_2-X_1)}{(1+X_1\beta_2)(1+X_2\beta_2)} & \frac{(X_2-X_1)\rho_*(X_1,X_2)}{(1+X_1\beta_2)(1+X_2\beta_2)} \\ \frac{X_1(X_1-X_2)}{(1+X_1\beta_2)(1+X_2\beta_2)} & \frac{X_1(X_2-X_1)\rho_*(X_1,X_2)}{(1+X_1\beta_2)(1+X_2\beta_2)} \end{array}\right]$$

is invertible or not. It is because

$$\mathbb{E}\left[\begin{array}{c} \frac{Y_{i2}-X_{i2}\beta_{1*}}{1+X_{i2}\beta_{2*}} - \frac{Y_{i1}-X_{i1}\beta_{1*}}{1+X_{i1}\beta_{2*}} \\ X_{i1}\left(\frac{Y_{i2}-X_{i2}\beta_{1*}}{1+X_{i2}\beta_{2*}} - \frac{Y_{i1}-X_{i1}\beta_{1*}}{1+X_{i1}\beta_{2*}}\right) \end{array}\right] - \mathbb{E}\left[\begin{array}{c} \frac{Y_2-X_2\beta_1}{1+X_2\beta_2} - \frac{Y_1-X_1\beta_1}{1+X_1\beta_2} \\ X_1\left(\frac{Y_2-X_2\beta_1}{1+X_2\beta_2} - \frac{Y_1-X_1\beta_1}{1+X_1\beta_2}\right) \end{array}\right] = A_2^{\text{pre}}(\beta)\left(\begin{array}{c} \beta_{1*}-\beta_1 \\ \beta_{2*}-\beta_2 \end{array}\right),$$

As in the previous section, we first discuss identification when $\text{Supp}(X_1, X_2) = \{a, b\}^2$, with each point having positive probability. In this case the determinant of $A_2^{\text{pre}}(\beta)$ equals

$$-(b-a)^3\left(\frac{f_{X_1,X_2}(a,b)}{(1+a\beta_2)(1+b\beta_2)}\right)^2 \frac{f_{X_1,X_2}(b,a)}{f_{X_1,X_2}(a,b)}[\rho_*(b,a)-\rho_*(a,b)],$$

which will be different from 0 if and only if $\rho_*(b,a) \neq \rho_*(a,b)$.

To investigate the possibility of point identification with binary X_t even when ρ_* is symmetric, we consider the case for which $T=3$ and $\text{Supp}(X_1, X_2, X_3) = \{a, b\}^3$ with $a \neq b$ and each triplet having positive probability. When $T=3$, the unconditional moments we have can be summarized by

$$\mathbb{E}\left[\tilde{X}_l\left(\frac{Y_t - X_t\beta_{1*}}{1+X_t\beta_{2*}} - \frac{Y_s - X_s\beta_{1*}}{1+X_s\beta_{2*}}\right)\right] = 0,$$

where $\tilde{X}_l$ denotes any measurable function of the constant and X^l with $l \le s < t$. We will focus on two of these conditions:

$$\mathbb{E}\left[\begin{array}{c} \frac{Y_2-X_2\beta_{1*}}{1+X_2\beta_{2*}} - \frac{Y_1-X_1\beta_{1*}}{1+X_1\beta_{2*}} \\ \frac{Y_3-X_3\beta_{1*}}{1+X_2\beta_{2*}} - \frac{Y_2-X_2\beta_{1*}}{1+X_2\beta_{2*}} \end{array}\right] = \left(\begin{array}{c} 0 \\ 0 \end{array}\right).$$

The β_* will be identified relative to any β such that the the expectations in the above equation evaluated at β are all well-defined if

$$\mathbb{E}\left[\begin{array}{cc} \frac{X_2-X_1}{(1+X_2\beta_{2*})(1+X_1\beta_{2*})} & \frac{(X_2-X_1)\theta_*(X_1,X_2,X_3)}{(1+X_2\beta_{2*})(1+X_1\beta_{2*})} \\ \frac{X_3-X_2}{(1+X_3\beta_{2*})(1+X_2\beta_{2*})} & \frac{(X_3-X_2)\theta_*(X_1,X_2,X_3)}{(1+X_3\beta_{2*})(1+X_2\beta_{2*})} \end{array}\right]$$

has rank 2. When the support of (X_1, X_2, X_3) equals $\{a, b\}^3$, the above matrix evaluated at β such that $(1 + a\beta_2)(1 + b\beta_2) \neq 0$ equals $\frac{b-a}{(1+a\beta_2)(1+b\beta_2)}$ times

$$\begin{bmatrix} f_{X_1X_2X_3}(a,b,a) - f_{X_1X_2X_3}(b,a,a) & \theta_*(a,b,a)f_{X_1X_2X_3}(a,b,a) - \theta_*(b,a,a)f_{X_1X_2X_3}(b,a,a) \\ f_{X_1X_2X_3}(a,a,b) - f_{X_1X_2X_3}(a,b,a) & f_{X_1X_2X_3}(a,a,b)\theta_*(a,a,b) - \theta_*(a,,b,a)f_{X_1X_2X_3}(a,b,a) \end{bmatrix}$$
$$+ \begin{bmatrix} f_{X_1X_2X_3}(a,b,b) - f_{X_1X_2X_3}(b,a,b) & \theta_*(a,b,b)f_{X_1X_2X_3}(a,b,b) - \theta_*(b,a,b)f_{X_1X_2X_3}(b,a,b) \\ f_{X_1X_2X_3}(b,a,b) - f_{X_1X_2X_3}(b,b,a) & \theta_*(b,a,b)f_{X_1X_2X_3}(b,a,b) - \theta_*(b,b,a)f_{X_1X_2X_3}(b,b,a) \end{bmatrix}.$$

Now suppose that $\theta_*(x_1, x_2, x_3) = \theta_*(\pi(x_1, x_2, x_3))$ for each permutation π, of (x_1, x_2, x_3). Then the matrix above simplifies to

$$\begin{bmatrix} q_1 & \theta_*(a,a,b)q_1 \\ q_2 & \theta_*(a,a,b)q_2 \end{bmatrix} + \begin{bmatrix} q_3 & \theta_*(a,b,b)q_3 \\ q_4 & \theta_*(a,b,b)q_4 \end{bmatrix},$$

where

$$\begin{aligned} q_1 &:= f_{X_1X_2X_3}(a,b,a) - f_{X_1X_2X_3}(b,a,a), \\ q_2 &:= f_{X_1X_2X_3}(a,a,b) - f_{X_1X_2X_3}(a,b,a), \\ q_3 &:= f_{X_1X_2X_3}(a,b,b) - f_{X_1X_2X_3}(b,a,b), \\ q_4 &:= f_{X_1X_2X_3}(b,a,b) - f_{X_1X_2X_3}(b,b,a). \end{aligned}$$

The determinant of this matrix equals $(\theta_*(a,a,b) - \theta_*(a,b,b))\,[q_2q_3 - q_1q_4]$. Thus, β_* will be identified if $(\theta_*(a,a,b) - \theta_*(a,b,b))$ and $[q_2q_3 - q_1q_4]$ are both different from 0. Note that a necessary condition for the latter is that (X_1, X_2, X_3) is not exchangeable. To further highlight the importance of non-exchangeability of $(X_1, X_2, ..., X_t)$ for identification we focus on $T = 2$ again:

Theorem 2 *Suppose $T = 2$, (X_1, X_2) is exchangeable and $\rho_*(x_1, x_2) = \rho_*(x_2, x_1)$ for almost every (x_1, x_2). Then β_* is not identified by the moment condition given in equation* (56).

Proof. (X_1, X_2) exchangeable means $f_{12}(x_1, x_2) = f_{12}(x_2, x_1)$ for almost every (x_1, x_2). Then

$$
\begin{aligned}
&\mathbb{E}\left[\frac{(X_2 - X_1)}{(1 + X_1\beta_2)(1 + X_2\beta_2)}\right] \\
&\quad = \int \frac{x_2}{(1 + x_1\beta_2)(1 + x_2\beta_2)} f_{12}(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 - \int \frac{x_1}{(1 + x_1\beta_2)(1 + x_2\beta_2)} f_{12}(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \\
&\quad = \int \frac{x_2}{(1 + x_1\beta_2)(1 + x_2\beta_2)} f_{12}(x_2, x_1)\mathrm{d}x_1\mathrm{d}x_2 - \int \frac{x_1}{(1 + x_1\beta_2)(1 + x_2\beta_2)} f_{12}(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \\
&\quad = 0, \\
&\mathbb{E}\left[\frac{(X_2 - X_1)\rho_*(X_1, X_2)}{(1 + X_1\beta_2)(1 + X_2\beta_2)}\right] \\
&\quad = \int \frac{x_2\rho_*(x_1, x_2)}{(1 + x_1\beta_2)(1 + x_2\beta_2)} f_{12}(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 - \int \frac{x_1\rho_*(x_1, x_2)}{(1 + x_1\beta_2)(1 + x_2\beta_2)} f_{12}(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \\
&\quad = \int \frac{x_2\rho_*(x_2, x_1)}{(1 + x_1\beta_2)(1 + x_2\beta_2)} f_{12}(x_2, x_1)\mathrm{d}x_1\mathrm{d}x_2 - \int \frac{x_1\rho_*(x_1, x_2)}{(1 + x_1\beta_2)(1 + x_2\beta_2)} f_{12}(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \\
&\quad = 0.
\end{aligned}
$$

Thus, the rank of $A_2^{\text{pre}}(\beta_2)$ is at most 1. ■

Our earlier results showed that when $T = 2$ and $\text{Supp}(X_1, X_2) = \{a, b\}^2$, β_* cannot be identified if ρ_* is symmetric, regardless of whether the joint density, $f_{X_1X_2}(x_1, x_2)$, of (X_1, X_2) is symmetric or not. That is because in that special case, symmetry of ρ_* means that ρ_* is constant and X_1 cannot be used as an instrumental variable for ρ_* in the second period equation because of the failure of the relevance condition. In contrast, Theorem 2 says that when $\text{Supp}(X_1, X_2)$ is richer, a sufficient condition for failure of identification is symmetry of both $f_{X_1X_2}$ and ρ_* in (x_1, x_2). When $\text{Supp}(X_1, X_2)$ is richer, Theorem 2 leaves open the possibility that β_* is identified even if ρ_* is symmetric, as long as $f_{X_1X_2}(x_1, x_2)$ is not symmetric, which is a testable condition. On the other hand, even when $f_{X_1X_2}(x_1, x_2)$ is not symmetric and $\text{Supp}(X_1, X_2)$ is richer than $\{a, b\}^2$, we cannot ensure that β_* is the only element of $\mathbb{R}^2$ that satisfies the moment conditions (56).

A.7 Data Appendix

We use detailed, student-level administrative records from the North Carolina Education Research Data Center (NCERDC). The records include information about all North Carolina public school students for the 2000-2013 school years from grades 3 through 10. The data contain an identifier for the student's school, reading and math test scores from standardized end-of-grade exams, an identifier for the teacher who monitored the standardized exams, and student characteristics including: sex, ethnicity, English proficiency, gifted/talented status, learning disabilities, and economic disadvantage.

To generate the analysis sample, we first select students following several criteria to help limit measurement error in exam scores. We do this by removing observations where the reported test score is from a retake exam or where a student is linked to multiple, different scores for a subject in an academic year. Additionally, since the scores come from end-of-grade standardized exams, we remove observations where the student is recorded as being in a grade that conflicts with the raw data file specific grade, since it is unclear which grade is accurate and relevant for the exam scores. Similarly, we remove observations where the student is either observed in multiple schools or associated with multiple teachers in a given grade and academic year since it is unclear which of these is accurate.

Our analysis to estimate teacher value added relies on the assumption that the observed teacher associated with the student level observation is the teacher who actually taught the student math. Since the teacher identified in the dataset is the teacher who proctored the exam, it is reasonably likely that this teacher also taught the student, but it is not guaranteed. We therefore further select on observed teacher characteristics to minimize the likelihood that the observed teacher is not the associated student's math teacher. First, we remove observations where the associated teacher is recorded as performing any administrative tasks, teaching nonstandard classes such as special education or honors courses, or teaching at a charter school. Second, we remove observations where the associated teachers are not recorded as specifically teaching both math and reading or as teaching a self-contained class. Third, we limit our sample of students to those in grades 3-5 (since beyond elementary school students are more likely to have multiple teachers in a given year), and keep observations where the associated teacher also only teaches in grades 3-5. In cases where teachers are recorded as teaching multiple grades, we identify the teacher's primary grade taught as the grade with the most students in it. We remove observations where the teacher has less than half of their students in their primary grade or the teacher teaches both grades 3 and 5. We also remove observations where the grade of the student does not match the teacher's primary grade taught. Finally, we restrict the sample to classrooms with more than 10 and no more than 24 valid student observations and where fewer than 50% of valid students are gifted/talented.

A.8 Estimating Equation

$$
\begin{aligned}
\dot{y}_{ijt} - \bar{\dot{y}}_{jt} = &\gamma^L(z^L_{it} - \bar{z}^L_{jt}) + \beta^L(\bar{\dot{y}}_{jt} z^L_{it} - \dot{y}_{ijt}\bar{z}^L_{jt}) \\
&+ \gamma^E(z^E_{it} - \bar{z}^E_{jt}) + \beta^E(\bar{\dot{y}}_{jt} z^E_{it} - \dot{y}_{ijt}\bar{z}^E_{jt}) \\
&+ \gamma^U(z^U_{it} - \bar{z}^U_{jt}) + \beta^U(\bar{\dot{y}}_{jt} z^U_{it} - \dot{y}_{ijt}\bar{z}^U_{jt}) \\
&+ \pi_1(\bar{z}^L_{jt} z^E_{it} - \bar{z}^E_{jt} z^L_{it}) \\
&+ \pi_2(\bar{z}^L_{jt} z^U_{it} - \bar{z}^U_{jt} z^L_{it}) \\
&+ \pi_3(\bar{z}^E_{jt} z^U_{it} - \bar{z}^U_{jt} z^E_{it}) + \epsilon_{ijt}
\end{aligned}
$$

where z_{it}^L is student i's lagged score, z_{it}^E is an indicator for economic disadvantage, and z_{it}^U is an indicator for underrepresented minority; $\bar{z}_{jt}$ are the respective classroom averages.

A.9 Appendix Tables

Table A1: Summary Statistics

	p5	p25	p50	Mean	SD	p75	p95
Math score	-1.58	-0.62	0.05	0.04	0.96	0.71	1.63
Reading score	-1.65	-0.61	0.07	0.03	0.97	0.75	1.53
Grade	4	4	4	4.47	0.50	5	5
Econ. disadv.	0	0	0	0.46	0.50	1	1
URM	0	0	0	0.35	0.48	1	1
Female	0	0	0	0.49	0.50	1	1
Asian	0	0	0	0.02	0.14	0	0
other race/ethnicity	0	0	0	0.04	0.21	0	0
Class size	12	15	18	17.85	3.48	21	23

N = 544,546 student-years

Table A2: Classroom Characteristics by Difference in Teacher Quality Estimates (Math)

	Teacher quality			Class average		
Teacher-years (N)	$\hat{\alpha}^{\text{Het.}} - \hat{\alpha}^{\text{No het.}}$	$\overline{\hat{\alpha}^{\text{Het.}}}$	$\overline{\hat{\alpha}^{\text{No het.}}}$	Lag score	Econ. dis.	URM
2611	< -0.02	-0.35	-0.32	0.01	0.58	0.53
5135	≥ -0.02 & < 0.01	-0.16	-0.15	-0.02	0.50	0.41
10130	≥ -0.01 & < 0	-0.02	-0.01	-0.03	0.46	0.33
8613	≤ 0.01 & > 0	0.13	0.13	0.03	0.44	0.32
3483	≤ 0.02 & > 0.01	0.27	0.25	0.15	0.44	0.35
1834	> 0.02	0.44	0.41	0.31	0.45	0.40
31806	Mean	0.03	0.03	0.03	0.47	0.36

Table A3: Classroom Characteristics by Difference in Teacher Quality Estimates (Reading)

	Teacher quality			Class average		
Teacher-years (N)	$\hat{\alpha}^{\text{Het.}} - \hat{\alpha}^{\text{No het.}}$	$\overline{\hat{\alpha}^{\text{Het.}}}$	$\overline{\hat{\alpha}^{\text{No het.}}}$	Lag score	Econ. dis.	URM
7101	< -0.02	-0.12	-0.08	0.03	0.50	0.39
4219	≥ -0.02 & < 0.01	-0.02	-0.01	-0.02	0.48	0.35
5113	≥ -0.01 & < 0	0.02	0.02	-0.02	0.47	0.34
4946	≤ 0.01 & > 0	0.06	0.05	-0.02	0.46	0.35
3866	≤ 0.02 & > 0.01	0.10	0.08	0.00	0.46	0.35
6561	> 0.02	0.21	0.17	0.12	0.44	0.37
31806	Mean	0.04	0.04	0.02	0.47	0.36

Table A4: AYP Difference-in-differences: # schools (# teachers) by subsample

	Control	Treated
2004	138 (1718)	101 (1612)
2005	118 (1063)	33 (188)
2006	76 (611)	51 (350)
2007	56 (509)	43 (336)
2008	53 (501)	23 (175)

Table A5: Impact of Accountability on Estimated Teacher Quality

	Math		Reading	
	No het.	Het.	No het.	Het.
Treat × Post	0.03*	0.03*	0.01	0.04**
	(0.01)	(0.02)	(0.01)	(0.02)

N =7,661 teacher-years (4,206 unique teachers). Standard errors clustered by school. All models include teacher and year fixed effects.