NBER WORKING PAPER SERIES

IDENTIFICATION OF NON-ADDITIVE FIXED EFFECTS MODELS: IS THE RETURN TO TEACHER QUALITY HOMOGENEOUS?

Jinyong Hahn
John D. Singleton
Neşe Yildiz

Working Paper 31384
http://www.nber.org/papers/w31384

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2023, Revised September 2025

We thank Bo Honoré for his helpful suggestion for extending the results under weak exogeneity. This Version: Thursday 11th September, 2025. Nese Yildiz has received financial support from the NSF through grant SES-1918985. Lena Harris provided excellent research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Identification of Non-Additive Fixed Effects Models: Is the Return to Teacher Quality Homogeneous?
Jinyong Hahn, John D. Singleton, and Neşe Yildiz
NBER Working Paper No. 31384
June 2023, Revised September 2025
JEL No. C12, C14, C31, C36, C52, H75, I21, I24

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 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 raised the effectiveness of teacher quality for lower performing students.

Jinyong Hahn
University of California, Los Angeles
hahn@econ.ucla.edu

John D. Singleton
University of Rochester
Department of Economics
and NBER
john.singleton@rochester.edu

Neşe Yildiz
University of Rochester
Department of Economics
nese.yildiz@rochester.edu

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.

Our results are obtained in "short" panels and they preserve key 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 the unobserved heterogeneity as well as for the other parameters in the model that govern average treatment effects, which can be continuous and multidimensional. We also extend the results to allow the parameters to depend on covariates. 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 increase in teacher quality is estimated to be about 25% *more* effective for a student one standard deviation below average in their prior score, all else equal. The estimates also indicate that the return to teacher quality is much lower for economically disadvantaged and underrepresented minority students. These findings have meaningful implications for estimates of individual teachers' qualities and for how equitably teacher quality is distributed.

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 teacher quality. In particular, we show that accountability pressure induced by failure to meet No Child Left Behind-era school-level Adequate Yearly Progress targets caused a meaningful increase in the effectiveness of teacher quality for lower achieving students (as captured by their prior score) in math. This finding implies that, all else equal, the net effect of accountability can be negative for high achieving students taught by high-quality teachers—and likewise for low achieving students taught by low-quality teachers. Accountability pressure also decreased the effectiveness of teacher quality for economically disadvantaged students.

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); Kim and Polachek (1994); Pesaran and Smith (1995); Durlauf et al. (2001); Browning and Carro (2007). These models are often called random coefficients models. Wooldridge (2003, 2005) provides conditions under which the linear fixed effect estimator consistently recovers treatment effects. The model we consider in this paper can be viewed as a random coefficient model where the intercept and the slope coefficient(s) of the treatment variable(s) depend on the same scalar unobserved heterogeneity. It can also be understood as a parametric approximation to the model in Evdokimov (2010), which likewise contains a non-additive scalar fixed effect.

The unique parsimonious structure of our model allows us to entertain a different set of identification conditions, based on model primitives, than prior work. An appealing feature of these conditions is that they clarify precisely the types of variation in observables needed to achieve identification; we show that estimation can be carried out via 2SLS after appropriately transforming the data. Arellano and Bonhomme (2012) focus on identifying the joint probability distribution of random coefficients, while Graham and Powell (2012) examine their expectation. Though these results also are obtained in "short" panels, they require restrictions on so-called stayers that we do not impose. Our model is likewise less general than the one in Evdokimov (2010), but that paper places restrictions on the conditional support of the fixed effects that we do not. However, because our insight is to leverage information about the unobserved heterogeneity embedded in other periods' treatment variables, our results do require an assumption that rules out independence between the fixed effects and the treatment variables.

The correlated random coefficient model considered in Verdier (2020), which restricts one random coefficient to be a linear function of another, is essentially the same as our base model and the identification strategies are closely related to our Proposition 1. That paper instead develops identification of ATE for stayers based on extrapolation (Lemieux, 1998; Suri, 2011), provided that treatment is binary. In contrast, our treatment variable X_{it} could be any type of random variable; it could even be a vector. Verdier (2020) also places restrictions on how controls enter the model that our results relax.[1]

Most of the literature focuses on the case of strictly exogeneity and it is sometimes not obvious

[1]See Remark 5.

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. A contribution of our paper is thus establishing identification results for random coefficient panel models with weakly 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 estimate individual teachers' test score value-added. This setup embodies an assumption—the return to teacher quality is the same for all students—that we show is rejected by the data. This connects with several recent papers that also allow for match effects but by instead positing that teachers have different skills at teaching different kinds of students (e.g. Ahn et al. 2020; Biasi et al. 2021; Bates et al. 2022; Delgado 2023). However, these papers either place restrictions on dimensionality (e.g. skills are two dimensional), assume the match effects are independent of student characteristics, and/or require data from across several years for the same teacher. In contrast, we identify match effects (under no distributional assumptions) arising from a scalar teacher quality that maps into learning in student- and context-specific ways. Our analysis of the effects of No Child Left Behind further shows that match effects depend on workplace characteristics, such as accountability incentives.[2]

The remainder of the paper is structured as follows. In Section 2, we introduce the model. Section 3 presents semi-parametric identification results under the assumption that the regressors are strictly exogeneous. In Section 4 we use a transformation to derive conditional moment restrictions. Using these moment conditions, we describe how our model parameters can be estimated using 2SLS. In Section 7, we then presents our empirical application to matched student-teacher data. Finally, Section 5 presents identification results when the regressors are only weakly exogenous or pre-determined, while Section 8 concludes. Proofs not presented in the main text are collected in the Appendix.

[2]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.

2 Model

This section presents the theoretical result on identification of the baseline model. The baseline model is parametrically specified and contains only one covariate which is the treatment variable. For notational ease, other covariates have been suppressed. All the results we present for the baseline model could be taken to hold conditional on these other covariates. In the next section, we discuss extensions of the model, including a nonparametric extension. This nonparametric extension should be of separate technical interest to theoretically interested readers.[3]

We first introduce the baseline model and discuss the parameters of interest. 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 between individuals; the second effect varies between individuals because X_t interacts with unobserved individual fixed effect. In particular, outcome Y_t is determined by

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

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 assume that the number of periods T is small and fixed. In fact, often we are going to assume $T = 2$. In the following, we will 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 appear in the evaluation of important marginal/treatment effects under the assumption the model (1) is the realization of a potential outcome model

$$Y_{it}(x) = \alpha_i + x\beta_{1*} + x\beta_{2*}\alpha_i + U_{it}.$$

[3]Freyberger (2018) analyzed a panel model where a nonparametric link function depends on an index consisting of time dummies interacted with fixed effects and the generic error term. Our own model does not allow such an index, and the two models do not nest each other, which leads to different mathematical treatments.

Under the additional assumption $\mathbb{E}(U_{it}) = 0$, the implied average treatment effects are given by

$$\mathbb{E}\big(Y_{it}(x') - Y_{it}(x)\big) = (x' - x)\beta_{1*} + (x' - x)\beta_{2*}\mathbb{E}\left(\alpha_i\right), \tag{2}$$

$$\mathbb{E}\big(Y_{it}(x') - Y_{it}(x)|X_{it} = x\big) = (x' - x)\beta_{1*} + (x' - x)\beta_{2*}\mathbb{E}\left(\alpha_i|X_{it} = x\right), \tag{3}$$

which explains the significance of the parameters in the context of treatment effect identification. In Appendix A.9, we discuss the importance of these parameters in some structural model that does not take the treatment effect formulation, thereby illustrating the potential structural interpretation of these parameters outside the treatment effects context.

Remark 1 *As is clear from (2) and (3), our specification imposes a restriction that the treatment effects are time-invariant. While this feature of the model is shared with many other papers in the random effects literature, which may not be plausible in some applications. In the next section, we study identification of a model in which* (β_{1*}, β_{2*}) *is a function of covariates* W_{it}*. Thus, in that model treatment effects are time-varying. It is possible to get identification results for other generalizations of the model that allow for the ATE to be time varying, although it is not discussed in the current paper.*

Remark 2 *It can be seen that the model (1) is observationally equivalent to*

$$Y_{it} = \tilde{\beta}_{0*} + \tilde{\alpha}_i + X_{it}\tilde{\beta}_{1*} + X_{it}\tilde{\beta}_{2*}\tilde{\alpha}_i + U_{it}, \tag{4}$$

where $\alpha_i = \tilde{\beta}_{0*} + \tilde{\alpha}_i$*,* $\beta_{1*} = \tilde{\beta}_{1*} - \tilde{\beta}_{2*}\tilde{\beta}_{0*}$ *and* $\tilde{\beta}_{2*} = \beta_{2*}$*. Assuming that it is based on the potential outcome model* $Y_{it}(x) = \tilde{\beta}_{0*} + \tilde{\alpha}_i + x\tilde{\beta}_{1*} + x\tilde{\beta}_{2*}\tilde{\alpha}_i + U_{it}$*, Under the additional assumption* $\mathbb{E}(U_{it}) = 0$*, which can be argued to be a normalization, the implied average treatment effects are given by* $\mathbb{E}\big(Y_{it}(x') - Y_{it}(x)\big) = (x' - x)\tilde{\beta}_{1*} + (x' - x)\tilde{\beta}_{2*}\mathbb{E}\left(\tilde{\alpha}_i\right)$*. Because* $\alpha_i = \tilde{\beta}_{0*} + \tilde{\alpha}_i$*,* $\beta_{1*} = \tilde{\beta}_{1*} - \tilde{\beta}_{2*}\tilde{\beta}_{0*}$ *and* $\tilde{\beta}_{2*} = \beta_{2*}$*, the two average treatment effects are identical. Therefore, identification of* β_{1*}*,* β_{2*}*,* $\mathbb{E}(\alpha_i)$ *and* $\mathbb{E}(\alpha_i|X_{it} = x)$ *in model (1) can be used even for the observationally equivalent model (4) for the evaluation of important marginal/treatment effects.*

3 Semiparametric Identification

We now present the identification results. For identification, we maintain $\mathbb{E}[|Y_{it}|] < \infty$, $\mathbb{E}[|X_{it}|] < \infty$, and $\mathbb{E}[|\alpha_i|] < \infty$. 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. Although this assumption is restrictive, it may be more believable in the context of fixed (short) T panels considered in this paper.[4] In Section A.7, we discuss identification of the baseline model under a pre-determinedness assumption instead.

Remark 3 *When paired with the structural model in equation (1), Assumption 1 more or less imposes a no-anticipation condition (so only treatment X_1 is relevant in period 1, and a no-carryover condition (so only treatment X_2 is relevant in period 2).*

Our identification approach will be based on first differencing. Specifically, under Assumption 1, for $s \neq t$ we have

$$\mathbb{E}[Y_t - Y_s|X_s = x_s, X_t = x_t] = (x_t - x_s)(\beta_{1*} + \beta_{2*}\mathbb{E}[\alpha|X_s = x_s, X_t = x_t]). \tag{5}$$

When $x_t \neq x_s$ we can divide both sides of equation (5) by $(x_t - x_s)$ to identify

$$\boxed{\beta_{1*} + \beta_{2*}\mathbb{E}[\alpha|X_s = x_s, X_t = x_t]}. \tag{6}$$

Note that equation (1) implies that

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

[4]Note that in Section 3.1, 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.

Subtracting x_t times the identified object (6) from (7), we identify

$$\boxed{\mathbb{E}\left[\alpha | X_s = x_s, X_t = x_t\right]}. \tag{8}$$

Inspection of (6) 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 (1) 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*

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

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,

$$\frac{\mathbb{E}\left[Y_2 - Y_1 | X_1 = b, X_2 = a\right]}{a - b} = \beta_{1*} + \beta_{2*}\rho_*(b, a), \tag{9}$$

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

$$\frac{\mathbb{E}\left[Y_2 - Y_1 | X_1 = c, X_2 = a\right]}{a - c} = \beta_{1*} + \beta_{2*}\rho_*(c, a), \tag{11}$$

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

Subtracting (11) from (9) 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 (10) and (12), 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 $\{a, b\}^2 \subseteq \text{Supp}(X_1, X_2)$ 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 $\rho_*(a, b) = \rho_*(b, a)$ is falsifiable.[5] The next proposition shows that even if

[5] If for example X_{it} is a binary variable taking values 0 and 1 only, then $\rho_*(1, 0) = \mathbb{E}(Y_2 | X_1 = 1, X_2 = 0)$ and $\rho_*(0, 1) = \mathbb{E}(Y_1 | X_1 = 0, X_2 = 1)$. In this case $\mathbb{E}(Y_2 | X_1 = d, X_2 = 1 - d), \mathbb{E}(Y_1 | X_1 = d, X_2 = 1 - d)$ for $d = 0, 1$ can be estimated at $\sqrt{n}$-rate using a linear regression. Thus, the null hypothesis that $\rho_*(1, 0) = \rho_*(0, 1)$ can be tested using a t-test.

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* $\text{Supp}(X_1, X_2, X_3) = \{a, b\}^3$ *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. ■

Remark 4 *The assumption that* $\mathbb{P}(X_{it}\beta_{2*} = -1) = 0$ *means that the fixed effect matters for the generation of the outcome for almost all treatment dose levels. In the dummy/binary treatment variable case, this assumption is violated when* $\beta_{2*} = -1$.[6] *In this case, note that* α_i *does not affect the potential treated outcome for any period, while it affects the potential untreated outcome for each period. Furthermore, since the identification of* (β_{1*}, β_{2*}) *does not rely on this assumption, this assumption is, in principle, falsifiable.*

3.1 Coefficients Depending on Covariates

In this section, we consider the following alternative model:

$$Y_t = \tilde{\alpha} + \beta_0(W_t) + X_t\tilde{\beta}_1(W_t) + X_t\beta_{2*}(W_t)\tilde{\alpha} + U_t. \tag{13}$$

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

[6] This assumption corresponds to $\alpha_1 = -1$ in Verdier (2020).

equation (13) we get

$$Y_t = \tilde{\alpha} + \beta_0(W_t) + \lambda_0(W_t) + X_t\tilde{\beta}_1(W_t) + X_t\beta_{2*}(W_t)\tilde{\alpha} + \varepsilon_t. \tag{14}$$

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

$$\begin{aligned} Y_t(x) &= \beta_0(W_t) + \lambda_0(W_t) + \tilde{\alpha} + x\tilde{\beta}_1(W_t) + x\beta_{2*}(W_t)\tilde{\alpha} + \varepsilon_t, \\ Y_t(x') &= \beta_0(W_t) + \lambda_0(W_t) + \tilde{\alpha} + x'\tilde{\beta}_1(W_t) + x'\beta_{2*}(W_t)\tilde{\alpha} + \varepsilon_t. \end{aligned}$$

Individual treatment effect equals

$$Y_t(x) - Y_t(x') = (x - x')\tilde{\beta}_1(W_t) + (x - x')\beta_{2*}(W_t)\tilde{\alpha}.$$

The average effect of treatment conditional on W_t is equal to

$$\mathbb{E}\left[Y_t(x) - Y_t(x')\middle| W_t\right] = (x - x')\tilde{\beta}_1(W_t) + (x - x')\beta_{2*}(W_t)\mathbb{E}\left[\tilde{\alpha}|W_t\right].$$

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 (14) can equivalently be written as

$$Y_t = \alpha + X_t\beta_{1*}(W_t) + X_t\beta_{2*}(W_t)\alpha + \varepsilon_t. \tag{15}$$

Remark 5 *In* (15)*, the coefficients β_1 and β_2 depend only on the contemporaneous value W_t of the covariates. We can in principle consider a more general specification, and allow them to depend on the entire history $(W_1, \ldots, W_T)$ of the covariates. Such a generalization would strictly nest the specification considered by (Verdier, 2020, Section 2.5), as long as time dummies are assumed to be zero. On the other hand, we expect the specification* (15) *is more likely to be adopted in practice.*

We now discuss the identification of the model (15). For this discussion, we modify our strict exogeneity assumption to the following.

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

To keep the notation simple, we assume that $\text{Supp}(W_t) = \text{Supp}(W_{t-1})$.[7] 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.

Theorem 2 *Suppose $T = 2$ and that Assumption 2 holds. Let*

$$\rho_*(x_1, x_2, w, w) = \mathbb{E}[\alpha|X_1 = x_1, X_2 = x_2, W_1 = W_2 = W],$$

and

$$\begin{aligned}\mathcal{A}_1^{Cov} := \{x_1 : \exists x_2, \tilde{x}_2 \text{ such that } (x_1, x_2), (x_1, \tilde{x}_2) \in \text{Supp}(X_1, X_2|W_1 = W_2 = w) \text{ with } x_1 \neq x_2,\\ x_1 \neq \tilde{x}_2, x_2 \neq \tilde{x}_2, \text{ and } \rho_*(x_1, w, x_2, w,) \neq \rho_*(x_1, w, \tilde{x}_2, w)\},\\ \mathcal{A}_2^{Cov} := \{x_2 : \exists x_1, \tilde{x}_1 \text{ such that } (x_1, x_2), (\tilde{x}_1, x_2) \in \text{Supp}(X_1, X_2|W_1 = W_2 = w) \text{ with } x_1 \neq x_2,\\ \tilde{x}_1 \neq x_2, x_1 \neq \tilde{x}_1, \text{ and } \rho_*(x_1, w, x_2, w) \neq \rho_*(\tilde{x}_1, w, x_2, w)\}.\end{aligned}$$

Suppose that $\mathcal{A}_1^{Cov} \cup \mathcal{A}_2^{Cov}$ is measurable and has strictly positive probability. Then $\beta_{1*}(w)$ and $\beta_{2*}(w)$ are identified.

Remark 6 *Based on Theorem 2, we can also identify $\mathbb{E}[\alpha]$. For this purpose, we write*

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

Then if the support of (X_t, W_t) is such that $1 + x\beta_2(w) = 0$ occurs with zero probability, $\mathbb{E}[\alpha|X_t = x, W_t = w]$ will be identified for almost every (x, w). Integrating X_t and W_t out we identify $\mathbb{E}[\alpha]$. Also, note that if X_{it} is continuously distributed, or if there exists a component, say W_{itj}, of W_{it} that is continuously distributed, such that and $\beta_{2*}(\cdot)$ is one-to-one and continuously differentiable in w_{ijt} for each $w_{ijt} \in \text{Supp}(W_{ijt})$, then $\mathbb{P}(X_{it}\beta_{2*}(W_{it}) = -1) = 0$.

Remark 7 *Propositions 1 and 2 as well as their proofs can be adopted for (15) in a similar*

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

fashion.

The identification approaches outlined so far suggest straightforward estimation strategies. When X_{it} or W_{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 estimated using this alternative approach.

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 5.

In the case of scalar X_{it}, the conditional moment restriction that forms the basis of estimation is presented in the following result:

Proposition 4 *Suppose Assumption* (1) *holds. Then*

$$\mathbb{E}\left[\left(Y_{i2} - Y_{i1}\right) - \left(X_{i2} - X_{i1}\right)\beta_{1*} - \left(X_{i2}Y_{i1} - X_{i1}Y_{i2}\right)\beta_{2*} \middle| X_{i1}, X_{i2}\right] = 0, \tag{16}$$

and

$$\mathbb{E}\left[g(X_{i1}, X_{i2})\left(\left(Y_{i2} - Y_{i1}\right) - \left(X_{i2} - X_{i1}\right)\beta_{1*} - \left(X_{i2}Y_{i1} - X_{i1}Y_{i2}\right)\beta_{2*}\right)\right] = 0 \tag{17}$$

for any measurable function $g(X_{i1}, X_{i2})$.

Proof. Because

$$Y_{it} - X_{it}\beta_{1*} = \alpha_i(1 + X_{it}\beta_{2*}) + U_{it}, \tag{18}$$

we have

$$(Y_{i2} - X_{i2}\beta_{1*})(1 + X_{i1}\beta_{2*}) - (Y_{i1} - X_{i1}\beta_{1*})(1 + X_{i2}\beta_{2*}) = U_{i2}(1 + X_{i1}\beta_{2*}) - U_{i1}(1 + X_{i2}\beta_{2*}),$$

which implies that

$$Y_{i2} - Y_{i1} - (X_{i2} - X_{i1})\beta_{1*} - (X_{i2}Y_{i1} - X_{i1}Y_{i2})\beta_{2*} = U_{i2}(1 + X_{i1}\beta_{2*}) - U_{i1}(1 + X_{i2}\beta_{2*}).$$

Strict exogeneity then implies that equation (16) holds, from which (17) follows by the law of iterated expectations. ■

Letting $z_i := (g_1(X_{i1}, X_{i2}), \ldots, g_m(X_{i1}, X_{i2}))^\top$ for some $1 \leq m < \infty$, it is straightforward to recognize that (17) implies consistency of the 2SLS estimator applied to the equation

$$Y_{i2} - Y_{i1} = \beta_1 (X_{i2} - X_{i1}) + \beta_2 (X_{i2}Y_{i1} - X_{i1}Y_{i2}) + \upsilon_i \tag{19}$$

using z_i as instruments for $X_{i2}Y_{i1} - X_{i1}Y_{i2}$. In particular, $z_i = X_{i2}X_{i1}$ is valid given the structural equation (1) and Assumption 1. Proposition 4 thus provides the basis for estimation of β_{1*} and β_{2*} via linear IV regression (i.e., 2SLS), which is what we adopted in the empirical application.[8] We formalize this by the corollary below.

Corollary 1 *Given an i.i.d. sample $(X_{i1}, X_{i2}, Y_{i1}, Y_{i2})$ satisfying Assumption 1, let $(\hat{\beta}_1, \hat{\beta}_2)$ denote the 2SLS estimates applied to* (19)*. Under standard assumptions*[9]*, $(\hat{\beta}_1, \hat{\beta}_2)$ is $\sqrt{n}$-consistent for (β_{1*}, β_{2*}).*

Remark 8 *The moment* (17) *holds for every measurable function g, which implies that the degree of overidentification can be argued to be infinity. One can therefore use the estimator due to Domínguez and Lobato (2004), although we imagine that the 2SLS with a finite number of instruments will be more often adopted in empirical practice.*

Remark 9 *If α is independent of (X_1, X_2), then functions of (X_1, X_2) will not satisfy the relevance condition. Since the relevance condition is empirically verifiable, failure of the relevance*

[8] $\mathbb{E}[\alpha_i | X_{it}]$ and $\mathbb{E}[\alpha_i]$ can then be estimated using equation (18) above, provided that $\mathbb{P}(1 + X_{it}\beta_{2*} \neq 0) = 1$ for some t. If $0 < \mathbb{P}(1 + X_{it}\beta_{2*} \neq 0) < 1$ for each t, then $\mathbb{E}[\alpha_i | X_{it}]$ is identified for $X_{it} \in \mathcal{I}_t$ with $\mathcal{I}_t := \{x \in \text{Supp}(X_{it}) : 1 + x\beta_{2*} \neq 0\}$.

[9] Standard assumptions include finite fourth moments for every regressor, error term, and the instruments, as well as the rank condition, i.e., the requirement that the matrix

$$\mathbb{E}\left[z_i \left[(X_{i2} - X_{i1}), (X_{i2}Y_{i1} - X_{i1}Y_{i2})\right]\right]$$

has full column rank.

condition may imply that α is in fact a random effect.

The result above easily generalizes the conditional moment restriction (16) to the case with vector-valued X_{it}:

Proposition 5 *Suppose Assumption 1 holds. Then*

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

Proof. Because $Y_{it} - X_{it}^\top \beta_{1*} = \alpha_i(1 + X_{it}^\top \beta_{2*}) + U_{it}$, we have

$$(Y_{i2} - X_{i2}^\top \beta_{1*})(1 + X_{i1}^\top \beta_{2*}) - (Y_{i1} - X_{i1}^\top \beta_{1*})(1 + X_{i2}^\top \beta_{2*}) = U_{i2}(1 + X_{i1}^\top \beta_{2*}) - U_{i1}(1 + X_{i2}^\top \beta_{2*}),$$

which implies that

$$Y_{i2} - Y_{i1} - (X_{i2} - X_{i1})^\top \beta_{1*} - \beta_{2*}^\top (X_{i1}X_{i2}^\top - X_{i2}X_{i1}^\top)\beta_{1*} - (X_{i2}Y_{i1} - X_{i1}Y_{i2})^\top \beta_{2*} \\ = U_{i2}(1 + X_{i1}^\top \beta_{2*}) - U_{i1}(1 + X_{i2}^\top \beta_{2*}).$$

The result then follows from Assumption 1. ■

Remark 10 *By the law of iterated expectations, the counterpart of* (17) *follows from Proposition 5.*

Remark 11 *This result is convenient because 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 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 3 *For each* $t = 1, 2, ..., T$, $\mathbb{E}\left[U_{it}|X_i^t\right] = \mathbb{E}\left[U_{it}|X_{it}\right] = 0$, *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 4 *For each* $t = 1, 2, ..., T$, $\mathbb{E}\left[\left|\frac{U_t}{1+X_2\beta_{2*}}\right|\right] < \infty$.

5.1 Conditional Moment Restrictions

Under Assumption 3, the identification arguments leading up to Theorem 1 no longer work. We can, however, still use the conditional moment condition.

$$\mathbb{E}\left[\frac{Y_{it} - X_{it}\beta_{1*}}{1 + X_{it}\beta_{2*}} - \frac{Y_{it-1} - X_{it-1}\beta_{1*}}{1 + X_{it-1}\beta_{2*}}\middle| X_i^{t-1}\right] = 0 \tag{20}$$

to identify the parameters.

Proposition 6 *Suppose that* $T = 2$, *and* $\{a, b\}^2 \subset \text{Supp}(X_1, X_2)$, *with each point having positive probability such that* $(1 + a\beta_{2*})(1 + b\beta_{2*}) \neq 0$. *Then* (β_{1*}, β_{2*}) *in* (20) *is identified if* $\rho_*(b, a) \neq \rho_*(a, b)$, *where* $\rho_*(x_1, x_2) := \mathbb{E}\left[\alpha|X_1 = x_2, X_2 = x_2\right]$.

When $T = 3$, we can exploit

$$0 = \mathbb{E}\left[\frac{Y_{i3} - X_{i3}\beta_{1*}}{1 + X_{i3}\beta_{2*}} - \frac{Y_{i2} - X_{i2}\beta_{1*}}{1 + X_{i2}\beta_{2*}}\middle| X_{i1}, X_{i2}\right] \tag{21}$$

as well.

Proposition 7 *Suppose that* $T = 3$, *and* $\{a, b\}^3 \subset \text{Supp}\left(X_{i1}, X_{i2}, X_{i3}\right)$ *with each point in this support associated with positive probability such that* $(1 + a\beta_{2*})(1 + b\beta_{2*}) \neq 0$. *Then,* (β_{1*}, β_{2*}) *in* (21) *is identified if* $\theta_*\left(a, b, a\right) - \theta_*\left(b, b, a\right) \neq 0$, *where* $\theta_*(x_1, x_2, x_3) := \mathbb{E}\left[\alpha|X_1 = x_2, X_2 = x_2, X_3 = x_3\right]$.

5.2 Unconditional Moment Restrictions 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:

$$\mathbb{E}\left[\begin{array}{c} \frac{Y_{i2}-X_{i2}\beta_{1*}}{1+X_{i2}\beta_{2*}} - \frac{Y_{i1}-X_{i1}\beta_{1*}}{1+X_{i1}\beta_{2*}} \\ X_{i1}\left(\frac{Y_{i2}-X_{i2}\beta_{1*}}{1+X_{i2}\beta_{2*}} - \frac{Y_{i1}-X_{i1}\beta_{1*}}{1+X_{i1}\beta_{2*}}\right) \end{array}\right] = \left(\begin{array}{c} 0 \\ 0 \end{array}\right). \tag{22}$$

Let $\beta := (\beta_1, \beta_2)^\top$.

Proposition 8 *Suppose that $T = 2$, and $\mathrm{Supp}(X_1, X_2) = \{a, b\}^2$, with each point having positive probability. Then $\beta_* = (\beta_{1*}, \beta_{2*})^\top$ in (22) is identified 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 $T = 3$ case. The unconditional moments then 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 $l \leq s < t$. We will focus on two of these conditions:

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

Proposition 9 *Suppose that $T = 3$ and $\mathrm{Supp}(X_1, X_2, X_3) = \{a, b\}^3$ with $a \neq b$ and each triplet having positive probability. Then β_* will be identified if and only if $(\theta_*(a, a, b) - \theta_*(a, b, b))$ and*

$[q_2 q_3 - q_1 q_4]$ *are both different from 0, where*

$$q_1 := f_{X_1X_2X_3}(a,b,a) - f_{X_1X_2X_3}(b,a,a),$$
$$q_2 := f_{X_1X_2X_3}(a,a,b) - f_{X_1X_2X_3}(a,b,a),$$
$$q_3 := f_{X_1X_2X_3}(a,b,b) - f_{X_1X_2X_3}(b,a,b),$$
$$q_4 := f_{X_1X_2X_3}(b,a,b) - f_{X_1X_2X_3}(b,b,a).$$

Remark 12 *Note that a necessary condition for* $(\theta_*(a,a,b) - \theta_*(a,b,b)) \neq 0$ *and* $[q_2 q_3 - q_1 q_4] \neq 0$ *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:

Proposition 10 *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* (22).

Remark 13 *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, Proposition 10 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, Proposition 10 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* (22).

The proofs of the propositions presented in this and previous subsection are provided in the Appendix. There we also provide an alternative identification strategy for the special case of binary treatment under weak exogeneity assumption.

6 Further Extensions

We extend our analysis in several directions in the appendix. In A.5, we consider the model of the form $Y_t = \alpha + X_t\beta_{11*} + X_t^2\beta_{12*} + X_t\beta_{21*}\alpha + X_t^2\beta_{22*}\alpha + U_t$, and establish how such a model can be identified. In A.6, we investigate identification if the unconditional moment restriction (17) is used as the sole basis of identification. In A.8, we consider identification of the transformation model of the form $h(Y_{it}, \delta_*) = \alpha_{*i} + X_{it}\beta_{1*} + X_{it}\beta_{2*}\alpha_{*i} + U_{*it}$.

7 Empirical Application: Heterogeneity in the Return to Teacher Quality

In this section, we apply our results to study heterogeneity in the return to teacher quality. 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; the value-added or quality of the teacher; and an error term. Considering a dataset drawn from a single school year, a representative setup is given by:

$$y_{js} = \gamma x_{js} + \alpha_j + \epsilon_{js} \tag{23}$$

where y_{js} is the test score of student s, who is taught by teacher j. Note that, rather from repeat observations over time, the panel structure arises in this setting from the grouping of multiple students (s) into a single classroom with one teacher (j); s plays the role of the t dimension in the previous sections. x_{js} represents observed covariates, such as include student s's sex, race, ethnicity, and (proxies for) economic advantage, as well as student s's score in their prior grade. α_j summarizes the quality of teacher j, while ϵ_{js} 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.[10]

[10] 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).

Our model allows for considering the case where education production is no longer additive in teacher 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 student attributes z_s, where $z_s \subseteq x_{js}$:

$$y_{js} = \gamma x_{js} + \alpha_j + \beta \alpha_j z_s + \upsilon_{sj} \tag{24}$$

This equation is equivalent to the standard setup above, except for the addition of the $\beta \alpha_j z_s$. This term captures the heterogeneity in the return to α_j, with β governing the magnitude and nature of the heterogeneity. For example, if z_s is s's prior score, students with higher prior scores will benefit relatively more from an increase in teacher quality when $\beta > 0$, while students with lower prior scores benefit more when $\beta < 0$ (all else held equal). In an Appendix, we show that equation (24) has a natural microfoundation in a model of endogenous match effects; the sign and magnitude of β depends on how student weights in the teacher's objective function and the teacher's effort costs depend on z_s.

7.1 Testing Homogeneity in the Return to Teacher Quality

We marshal student-teacher matched administrative data from North Carolina to test whether the return to return to teacher quality is the same for all students. 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.

The equation we take to the data pools observations from both grades and across all years:

$$y_{jst} = \gamma x_{jst} + \alpha_{jt} + \beta \alpha_{jt} z_{st} + \upsilon_{sjt} \tag{25}$$

In pooling data from multiple years, we add year t subscripts to variables, but the fundamental panel structure remains classrooms. For both subjects (math and reading), we estimate two specifications: one where z_{st} includes only student s's prior score at time t and a richer model

that additionally incudes indicators for economic disadvantage and underrepresented minority (i.e. Black or Hispanic) status in z_{st}.[11] x_{st} includes the elements in z_{st} as well as lagged score squared and cubed, 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 (25), we first residualize y_{jst} with respect to the covariates not in z_{st} (i.e. those that do not interact with teacher quality).[12] Denote by $\dot{y}_{jst}$ the residualized test score. We then apply the transformation from the previous section with respect to the classroom average, $\bar{\dot{y}}_{jt}$.[13] In the case where z_{st} contains one variable (e.g. only lagged score), we obtain the following conditional moment restriction similar to the conditional moment restriction (16) discussed in Section 4:

$$E[(\dot{y}_{jst} - \bar{\dot{y}}_{jt}) - \gamma(z_{st} - \bar{z}_{jt}) - \beta(\bar{\dot{y}}_{jt} z_{st} - \dot{y}_{jst}\bar{z}_{jt})|x_{st}, \bar{x}_{jt}] = 0 \tag{26}$$

where $\bar{z}_{jt}$ collects classroom average characteristics among teacher j's students. This expectation forms the basis of the estimator we use—2SLS. $\bar{\dot{y}}_{jt} z_{st} - \dot{y}_{jst}\bar{z}_{jt}$ is an endogenous regressor. The instruments we use for our main results are the other elements of x_{st} not in z_{st}, e.g. whether female; class averages of the covariates $\bar{x}_{jt}$; and student-class average interactions, e.g. $x_{st}\bar{x}_{jt}$. For models that include economic disadvantage and underrepresented minority status in z_{st}, the estimating equation includes several additional terms.[14] We cluster standard errors by teacher.

Table 1 presents the results. The results allowing for heterogeneity in the return to teacher quality are presented side-by-side with results from estimating models that assume homogeneity. While columns (2) and (5) for math and reading, respectively, do not find evidence for heterogeneity when restricting it to only along the dimension of lagged score, columns (3) and (6),

[11] We focus on these characteristics because they are central to education policy debates about distributional effects, so heterogeneous returns along them are of primary substantive interest, because (unlike most other attributes of students in the data) they exhibit sufficient within-classroom variation to deliver precise interaction estimates.

[12] Specifically, we regress test scores on the covariates not in z_{st}, teacher fixed effects, and teacher-specific slopes on z_{st} and subtract out the effects of the controls.

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

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

which allow for heterogeneity along three student dimensions, clearly show that the assumption of common returns to teacher quality is rejected by the data. For example, the interaction on lagged score and teacher quality in column (3) is -0.25 and statistically significant. The economic interpretation of this is that a given increase in teacher quality would be 25% more effective for a student who is one standard deviation below average in their prior score. The estimates, which are very similar for math and reading, also show that an increase in teacher quality is about 70% less effective for an economically disadvantaged student and 50% less effective for an underrepresented minority student.

Table 1: Estimates of Education Production Function

	Math score			Reading score		
	(1)	(2)	(3)	(4)	(5)	(6)
Lagged score	0.82***	0.83***	0.84***	0.79***	0.79***	0.81***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Lagged score × Teacher quality		-0.04	-0.25***		-0.01	-0.23***
		(0.03)	(0.02)		(0.02)	(0.02)
Econ. disadv.	-0.07***	-0.07***	-0.05***	-0.09***	-0.09***	-0.07***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Econ. disadv. × Teacher quality			-0.73***			-0.75***
			(0.05)			(0.04)
URM	-0.06***	-0.06***	-0.05***	-0.09***	-0.09***	-0.07***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
URM × Teacher quality			-0.47***			-0.53***
			(0.04)			(0.03)
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)

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.

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 (in both subjects). These results have important implications for estimates of individual teachers' qualities and for how teacher quality is distributed across students. Appendix Table A2 shows that teacher quality in math is underestimated by the homogeneous model for those serving the most economically advantaged quintile of

classrooms, but that teacher quality in reading is underestimated (by 3% of a test score standard deviation on average) for those serving the least advantaged quintile.

7.2 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—as predicted by the model of teacher effort we sketch in the Appendix. 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).[15]

To answer this question, a naive approach might be to estimate the "production function" equation (25) separately for schools under accountability pressure and those not, obtaining two estimates of β which could then be compared. A worry with that comparison, however, would be that any difference obtained might reflect unobserved differences between schools rather than effects of accountability alone. Similarly, one could compare estimates of β from before and after the *same* school fails to meet AYP, but this could conflate time trends with the policy effect.

The approach we take instead purges unobserved heterogeneity and secular effects from the comparison of interest under a parallel trends assumption: in the absence of accountability pressure, the change in the production function between adjacent years for schools that enter into accountability pressure would have been the same as the (contemporaneous) change for schools that do not. Under this assumption, we can generate four estimates of β for any given year—for those schools under accountability pressure (β_1^T) and those not (β_1^C) and for the same two groups of schools the prior year (β_0^T and β_0^C)—and then difference to answer how the return to teacher quality changes due to accountability pressure:

$$\Delta\beta = (\beta_1^T - \beta_0^T) - (\beta_1^C - \beta_0^C)$$

We implement this idea in a way that uses the variation in the timing of accountability

[15]These data were generously provided by Josh Hollinger, who collected them from the North Carolina Department of Public Instruction's webpage.

pressure across several years but that is robust to issues with negative weights in contexts with staggered adoption (Goodman-Bacon, 2021; Roth et al., 2023). Specifically, we build our estimation sample by stacking subsamples indexed by year. For each year τ, we code as treated those schools that missed AYP for the first time the prior year and so experience accountability pressure that year. We code and keep as controls all those schools who have not yet experienced accountability pressure at τ—but who eventually will. Schools that previously failed AYP are excluded from subsequent subsamples and thus do not re-enter the sample as contaminated controls.[16] 2004 was the first year a school could be pressured, while schools treated in 2009 have no valid controls since the program ended that year. The variation is summarized in Table A3. The subsamples are then pooled in estimation and weighted by their inverse number of observations, meaning each cohort contributes the same (positive) weight.

The estimating equation we take to the data is given by:[17]

$$\begin{aligned}\dot{y}_{jst} - \bar{\dot{y}}_{jt} =& \gamma_0(z_{st} - \bar{z}_{jt}) + \beta_0(\bar{\dot{y}}_{jt} z_{st} - \dot{y}_{jst}\bar{z}_{jt}) + ... + \epsilon^{\tau}_{jst} \\ &+ \mathbf{1}[t \geq \tau][\gamma_1(z_{st} - \bar{z}_{jt}) + \beta_1(\bar{\dot{y}}_{jt} z_{st} - \dot{y}_{jst}\bar{z}_{jt})] \\ &+ Treat^{\tau}_{j}[\gamma_2(z_{st} - \bar{z}_{jt}) + \beta_2(\bar{\dot{y}}_{jt} z_{st} - \dot{y}_{jst}\bar{z}_{jt})] \\ &+ Treat^{\tau}_{j}\mathbf{1}[t \geq \tau][\gamma_3(z_{st} - \bar{z}_{jt}) + \beta_3(\bar{\dot{y}}_{jt} z_{st} - \dot{y}_{jst}\bar{z}_{jt})] \end{aligned} \tag{27}$$

where $Treat^{\tau}_{j}$ is an indicator for whether teacher j was treated in year τ.[18] Note that instead of estimating four sets of production function parameters and then appropriately differencing, our estimating equation makes the changes in the parameters associated with the onset of accountability pressure the target estimands: β_3 captures the effect of accountability on the return to teacher quality, i.e. $\Delta\beta$

Table 2 reports estimates of how accountability pressure impacts the return to teacher quality. For math, accountability pressure makes the interaction of a student's lagged score with teacher

[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). We cluster standard errors by school to accommodate this duplication.

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

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

quality more negative, meaning that accountability pressure increases the relative return to teacher quality for lower ability students. The point estimate is -0.40—implying that a given increase in teacher quality is 40 points more effective for a student one standard deviation below average than absent accountability pressure—and is statistically different from zero. A given increase in teacher quality is also nearly 60 points less effective for an economically disadvantaged student due to accountability pressure. These results for math provide novel evidence that NCLB-era accountability raised the lowered return to teacher quality for higher ability student and for economically disadvantaged students. We do not find evidence that accountability pressure affects returns to teacher quality in reading.

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

	Math	Reading
Lagged score × Teacher quality × Treat × Post	-0.40***	-0.02
	(0.10)	(0.12)
Econ. disadv. × Teacher quality × Treat × Post	-0.58***	0.00
	(0.20)	(0.17)
URM × Teacher quality × Treat × Post	0.09	0.00
	(0.19)	(0.19)

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

8 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, though it also reduced effectiveness of teaching quality for economically disadvantaged students. These findings suggest several avenues for future work, including extending the model to incorporate a richer set of covariates—such as classroom, school, or neighborhood resources—that may vary systematically with student disadvantage and help isolate the mechanisms underlying heterogeneity in teacher effectiveness.

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, and J. James (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.

Bates, M. D., M. Dinerstein, A. C. Johnston, and I. Sorkin (2022). Teacher labor market equilibrium and the distribution of student achievement. Technical report, National Bureau of Economic Research.

Biasi, B., C. Fu, and J. Stromme (2021). Equilibrium in the market for public school teachers: District wage strategies and teacher comparative advantage. Technical report, National Bureau of Economic Research.

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.

Delgado, W. (2023). Disparate teacher effects, comparative advantage, and match quality.

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.

Domínguez, M. A. and I. N. Lobato (2004). Consistent estimation of models defined by conditional moment restrictions. *Econometrica 72*(5), 1601–1615.

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.

Freyberger, J. (2018). Non-parametric panel data models with interactive fixed effects. *The Review of Economic Studies 85*(3), 1824–1851.

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

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of econometrics 225*(2), 254–277.

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 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.

Kim, M.-K. and S. W. Polachek (1994). Panel estimates of the gender earnings gap: individual-specific intercept and individual-specific slope models. *Journal of Econometrics 61*(1), 23–42.

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

Lemieux, T. (1998). Estimating the effects of unions on wage inequality in a panel data model with comparative advantage and nonrandom selection. *Journal of Labor Economics 16*(2), 261–291.

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.

Roth, J., P. H. Sant'Anna, A. Bilinski, and J. Poe (2023). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics 235*(2), 2218–2244.

Suri, T. (2011). Selection and comparative advantage in technology adoption. *Econometrica 79*(1), 159–209.

Verdier, V. (2020). Average treatment effects for stayers with correlated random coefficient models of panel data. *Journal of Applied Econometrics 35*(7), 917–939.

Wooldridge, J. M. (2003). 03.2. 1. fixed effects estimation of the population-averaged slopes in a panel data random coefficient model. *Econometric Theory 19*(2), 411–412.

Wooldridge, J. M. (2005). Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *The Review of Economics and Statistics 87*(2), 385–390.

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

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

$$\frac{\mathbb{E}\left[Y_2 - Y_1 | X_1 = a, X_2 = b\right]}{b-a} = \beta_{1*} + \beta_{2*}\rho_*(a,b), \tag{29}$$

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 (28) or (29). ■

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 $\theta_*(a,a,b) \neq \theta_*(a,b,b)$.

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

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

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 Proof of Theorem 2

Proof. The proof of this theorem is almost the same of that of Theorem 1; equations (9) - (12) should be replaced with the corresponding versions in which the event being conditioned on should be changed to $\{X_1 = x_1, W_1 = w, X_2 = x_2, W_2 = w\}$ in all the conditional expectations for each $(x_1, x_2) \in \mathcal{A}_1^{Cov} \cup \mathcal{A}_2^{Cov}$. ■

A.5 Higher Order Terms

Consider

$$Y_t = \alpha + X_t\beta_{11*} + X_t^2\beta_{12*} + X_t\beta_{21*}\alpha + X_t^2\beta_{22*}\alpha + U_t. \tag{30}$$

We maintain the strict exogeneity assumption (Assumption 1).

Proposition 11 *Suppose that $T = 2$, and that there exist x_1, x_2, x_3, x_4 all distinct from each other and distinct from 0 such that $(0, x_j) \in \text{Supp}(X_1, X_2)$ for $j \in \{1, 2, 3, 4\}$. Then if the matrix*

$$\begin{bmatrix} 1 & \rho_*(0,x_1) & x_1 & x_1\rho_*(0,x_1) \\ 1 & \rho_*(0,x_2) & x_2 & x_2\rho_*(0,x_2) \\ 1 & \rho_*(0,x_3) & x_3 & x_3\rho_*(0,x_3) \\ 1 & \rho_*(0,x_4) & x_4 & x_4\rho_*(0,x_4) \end{bmatrix} \tag{31}$$

is invertible, $\beta_* = (\beta_{11*}, \beta_{12*}, \beta_{21*}, \beta_{22*})$ in the model (30) is identified.

Proof. 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 $(0, x_j) \in \text{Supp}(X_1, X_2)$ for $j \in \{1, 2, 3, 4\}$. With this assumption, $\mathbb{E}[Y_1|X_1 = 0, X_2 = x_j] = \rho_*(0, x_j)$ for $j = 1, 2, 3, 4$. So for each x such that $(0, x) \in \text{Supp}(X_1, X_2)$, $\rho_*(0, x)$ is identified. Then for $x \neq 0$, we have

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

In other words, we have

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

which implies that the invertibility of the matrix (31) implies identification. Note that since $\rho_*(0, x_j)$ is identified, the invertibility of this matrix is verifiable. ■

Remark 14 *The proof of Proposition 11 indicates that 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 (30) without requiring that the support of X_1 contains 0. For this purpose, we impose the condition that $T = 3$.

Proposition 12 *Suppose that $T = 3$, 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. Then $\beta_* = (\beta_{11*}, \beta_{12*}, \beta_{21*}, \beta_{22*})$ in the model (30) is identified.*

Proof. Note that

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

Taking differences and rearranging we get

$$\begin{aligned}&\frac{\mathbb{E}\left[Y_2|X_1 = x, X_2 = x_j, X_3 = x_k\right] - \mathbb{E}\left[Y_1|X_1 = x, X_2 = x_j, X_3 = x_k\right]}{x_j - x}\\ &\qquad = \beta_{11*} + \theta_*(x, x_j, x_k)\beta_{21*} + (x_j + x)\left(\beta_{12*} + \theta_*(x, x_j, x_k)\beta_{22*}\right),\end{aligned} \tag{32}$$

and

$$\begin{aligned}&\frac{\mathbb{E}\left[Y_3|X_1 = x, X_2 = x_j, X_3 = x_k\right] - \mathbb{E}\left[Y_1|X_1 = x, X_2 = x_j, X_3 = x_k\right]}{x_k - x}\\ &\qquad = \beta_{11*} + \theta_*(x, x_j, x_k)\beta_{21*} + (x_k + x)\left(\beta_{12*} + \theta_*(x, x_j, x_k)\beta_{22*}\right).\end{aligned} \tag{33}$$

Differencing once more yields

$$\frac{\frac{\mathbb{E}[Y_2|X_1=x,X_2=x_j,X_3=x_k]-\mathbb{E}[Y_1|X_1=x,X_2=x_j,X_3=x_k]}{x_j-x} - \frac{\mathbb{E}[Y_3|X_1=x,X_2=x_j,X_3=x_k]-\mathbb{E}[Y_3|X_1=x,X_2=x_j,X_3=x_k]}{x_k-x}}{x_j - x_k} = \beta_{12*} + \theta_*(x, x_j, x_k)\beta_{22*}. \tag{34}$$

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

$$\boxed{\beta_{12*} + \theta_*(x, x_j, x_k)\beta_{22*}}. \tag{35}$$

Plugging (35) into (32) or (33) identifies

$$\boxed{\beta_{11*} + \theta_*(x, x_j, x_k)\beta_{12*}}. \tag{36}$$

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 equa-

tion $\mathbb{E}[Y_1|X_1 = x, X_2 = x_j, X_3 = x_k]$ identifies

$$\boxed{\theta_*(x, x_j, x_k)}. \tag{37}$$

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

$$\boxed{\beta_{12*} + \theta_*(x, x_j, x_l)\beta_{22*}}, \quad \boxed{\beta_{11*} + \theta_*(x, x_j, x_l)\beta_{21*}}, \quad \boxed{\theta_*(x, x_j, x_l)}. \tag{38}$$

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 (35), (36), (37), and (38). ■

A.6 Conditional and Unconditional Moment Restrictions Under Strict Exogeneity

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 exogeneity assumption.

A.6.1 Conditional Moment Restriction with $T = 2$

When $T = 2$, we have the conditional moment restriction

$$\mathbb{E}\left[\left(Y_{i2} - Y_{i1}\right) - \left(X_{i2} - X_{i1}\right)\beta_{1*} - \left(X_{i2}Y_{i1} - X_{i1}Y_{i2}\right)\beta_{2*} \middle| X_{i1}, X_{i2}\right] = 0, \tag{39}$$

which delivers the estimating equation

$$\mathbb{E}\left[\left(Y_{i2} - Y_{i1}\right) - \left(X_{i2} - X_{i1}\right)\beta_1 - \left(X_{i2}Y_{i1} - X_{i1}Y_{i2}\right)\beta_2 \middle| X_{i1}, X_{i2}\right] = 0. \tag{40}$$

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

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

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

$$\begin{bmatrix} x_2 - x_1 & \left(x_2 - x_1\right)\rho_*\left(x_1, x_2\right) \\ x_2' - x_1' & \left(x_2' - x_1'\right)\rho_*\left(x_1', x_2'\right) \end{bmatrix}$$

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

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

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

A.6.2 Unconditional Moment Restriction with $T = 2$

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.

A.6.3 Conditional Moment Restriction with $T = 3$

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

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

which delivers the estimating equation

$$\mathbb{E}\left[(Y_t - Y_s) - (X_t - X_s)\beta_1 - (X_tY_s - X_sY_t)\beta_2\middle| X_1, X_2, X_3\right] = 0. \tag{41}$$

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

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

In particular, the identification would be achieved if there exist $(X_1, X_2, X_3) = (x_1, x_2, x_3), (x_1', x_2', x_3')$ such that the matrix

$$\left[\begin{array}{cc} x_3 - x_2 & (x_3 - x_2)\theta_*(x_1, x_2, x_3) \\ x_2' - x_1' & (x_2' - x_1')\theta_*(x_1', x_2', x_3') \end{array}\right]$$

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 $\theta_*(a,b,b) \neq \theta_*(a,a,b)$.

A.6.4 Unconditional Moment Restriction with $T = 2$

We can also derive unconditional moment restrictions, we can derive a unconditional moment restriction from (41). For example, we can use

$$\mathbb{E}\begin{bmatrix} X_1\{(Y_2-Y_1)-(X_2-X_1)\beta_{1*}-(X_2Y_1-X_1Y_2)\beta_{2*}\} \\ X_2\{(Y_2-Y_1)-(X_2-X_1)\beta_{1*}-(X_2Y_1-X_1Y_2)\beta_{2*}\} \\ X_3\{(Y_2-Y_1)-(X_2-X_1)\beta_{1*}-(X_2Y_1-X_1Y_2)\beta_{2*}\} \\ X_1\{(Y_3-Y_2)-(X_3-X_2)\beta_{1*}-(X_3Y_2-X_2Y_3)\beta_{2*}\} \\ X_2\{(Y_3-Y_2)-(X_3-X_2)\beta_{1*}-(X_3Y_2-X_2Y_3)\beta_{2*}\} \\ X_3\{(Y_3-Y_2)-(X_3-X_2)\beta_{1*}-(X_3Y_2-X_2Y_3)\beta_{2*}\} \end{bmatrix} = 0$$

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

$$\mathbb{E}\begin{bmatrix} X_1(X_2-X_1) & X_1(X_2Y_1-X_1Y_2) \\ X_2(X_2-X_1) & X_2(X_2Y_1-X_1Y_2) \\ X_3(X_2-X_1) & X_3(X_2Y_1-X_1Y_2) \\ X_1(X_3-X_2) & X_1(X_3Y_2-X_2Y_3) \\ X_2(X_3-X_2) & X_2(X_3Y_2-X_2Y_3) \\ X_3(X_3-X_2) & X_3(X_3Y_2-X_2Y_3) \end{bmatrix}$$

is equal to 2 or not.

A.7 Identification Under Weak Exogeneity

In this section we present the proofs of propositions stated in section (5). We also discuss an alternative identification strategy under weak exogeneity assumption for the special case of binary tretment.

A.7.1 Proof of Proposition 6

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

$$0 = \mathbb{E}\left[\frac{(X_{i1}-X_{i2})\,\alpha_i}{(1+X_{i2}\beta_2)(1+X_{i1}\beta_2)}(\beta_{2*}-\beta_2) + \frac{X_{i1}-X_{i2}}{(1+X_{i2}\beta_2)(1+X_{i1}\beta_2)}(\beta_{1*}-\beta_1)\middle| X_{i1}\right] \quad (42)$$

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. It is straightforward to show that requirement (42) becomes equivalent to $(\beta_1, \beta_2) = (\beta_{1*}, \beta_{2*})$ be the only solution to

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

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

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

■

A.7.2 Proof of Proposition 7

Proof. 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*}). 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

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

■

Remark 15 *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). 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.*

Before we end this section, let us consider the $T = 2$ case again. Suppose there exists $E \subseteq \text{Supp}(X_1)$ 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 (42) 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[19]

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

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.

A.7.3 Proof of Proposition 8

Proof. Whether $\beta_* = (\beta_{1*}, \beta_{2*})^\top$ is identified relative to β by the moment conditions given above depends on whether

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

is invertible or not. It is because

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

As in the previous section, we first discuss identification when $\text{Supp}(X_1, X_2) = \{a, b\}^2$, with

[19] In writing these equations we assumed that for each $x_1 \in E$ (42) is well-defined.

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 $\rho_*(b,a)\neq\rho_*(a,b)$. ■

A.7.4 Proof of Proposition 9

Proof. 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

$$\left[\begin{array}{cc}f_{X_1X_2X_3}(a,b,a)-f_{X_1X_2X_3}(b,a,a) & \theta_*(a,b,a)f_{X_1X_2X_3}(a,b,a)-\theta_*(b,a,a)f_{X_1X_2X_3}(b,a,a)\\ f_{X_1X_2X_3}(a,a,b)-f_{X_1X_2X_3}(a,b,a) & f_{X_1X_2X_3}(a,a,b)\theta_*(a,a,b)-\theta_*(a,,b,a)f_{X_1X_2X_3}(a,b,a)\end{array}\right]$$
$$+\left[\begin{array}{cc}f_{X_1X_2X_3}(a,b,b)-f_{X_1X_2X_3}(b,a,b) & \theta_*(a,b,b)f_{X_1X_2X_3}(a,b,b)-\theta_*(b,a,b)f_{X_1X_2X_3}(b,a,b)\\ f_{X_1X_2X_3}(b,a,b)-f_{X_1X_2X_3}(b,b,a) & \theta_*(b,a,b)f_{X_1X_2X_3}(b,a,b)-\theta_*(b,b,a)f_{X_1X_2X_3}(b,b,a)\end{array}\right].$$

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

$$\left[\begin{array}{cc}q_1 & \theta_*(a,a,b)q_1\\ q_2 & \theta_*(a,a,b)q_2\end{array}\right]+\left[\begin{array}{cc}q_3 & \theta_*(a,b,b)q_3\\ q_4 & \theta_*(a,b,b)q_4\end{array}\right].$$

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. ■

A.7.5 Proof of Proposition 10

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 $A_2^{\mathrm{pre}}(\beta_2)$ is at most 1. ■

A.7.6 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.

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

Therefore,

$$\rho_*(a, b) + a[\beta_{1*} + \beta_{2*}\rho_*(a, b)] = \frac{\mathbb{E}\left[Y_1|X_1 = a\right] - \mathbb{E}\left[Y_2|X_1 = X_2 = a\right]\mathbb{P}(X_2 = a|X_1 = a)}{\mathbb{P}(X_2 = b|X_1 = a)}. \quad (43)$$

Moreover,

$$\mathbb{E}\left[Y_2|X_1=a,X_2=b\right]=\rho_*(a,b)+b\left(\beta_{1*}+\beta_{2*}\rho_*(a,b)\right). \tag{44}$$

Combining (43) and (44), we can see that $\boxed{\beta_{1*}+\beta_{2*}\rho_*(a,b)}$ is identified by

$$\frac{\mathbb{E}\left[Y_2|X_1=a,X_2=b\right]-\frac{\mathbb{E}[Y_1|X_1=a]-\mathbb{E}[Y_2|X_1=X_2=a]\mathbb{P}(X_2=a|X_1=a)}{\mathbb{P}(X_2=b|X_1=a)}}{b-a}=\beta_{1*}+\beta_{2*}\rho_*(a,b). \tag{45}$$

Next, by subtracting b times (45) from (44), 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,

$$\begin{aligned}
&\frac{\mathbb{E}\left[Y_2|X_1=X_2=a\right]-\mathbb{E}\left[Y_3|X_1=X_2=X_3=a\right]\mathbb{P}(X_3=a|X_1=X_2=a)}{\mathbb{P}(X_2=b|X_1=X_2=a)}\\
&\quad=\theta_*(a,a,b)+a\left(\beta_{1*}+\beta_{2*}\theta_*(a,a,b)\right).
\end{aligned} \tag{46}$$

We also have

$$\mathbb{E}\left[Y_3|X_1=X_2=a,X_3=b\right]=\theta(a,a,b)+b\left(\beta_{1*}+\beta_{2*}\theta_*(a,a,b)\right). \tag{47}$$

From (46) and (47), we identify

$$\boxed{\beta_{1*}+\beta_{2*}\theta_*(a,a,b)},\quad\text{and}\quad\boxed{\theta_*(a,a,b)}. \tag{48}$$

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

$$\boxed{\beta_{1*}+\beta_{2*}\theta_*(b,b,a)},\quad\text{and}\quad\boxed{\theta_*(b,b,a)}. \tag{49}$$

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

(48) and (49).

A.8 Extension to Transformation Models

In this section, we discuss some extensions of our model. The first extension we consider is a parametric transformation model:

$$h\left(Y_{it}, \delta_*\right) = \alpha_{*i} + X_{it}\beta_{1*} + X_{it}\beta_{2*}\alpha_{*i} + U_{*it},$$

where the function h is known up to a finite dimensional parameter. We note that the moment conditions we discuss under strict and weak exogeneity immediately apply to this case.

Next, we consider a non-parametric transformation model. We achieve this by considering the identification of a model expressed in terms of m_*:

$$Y_{it} = m_*(\alpha_{*i} + X_{it}\beta_{1*} + X_{it}\beta_{2*}\alpha_* + U_{*it}), \tag{50}$$

with a strictly increasing and continuously differentiable unknown function m_*. We discuss partial identification of this model. The tools we use in this discussion are similar to those used in Evdokimov (2010), which studies $Y_{it} = m(X_{it}, \alpha_i) + U_{it}$, with m strictly monotone in α_i. He studies point identification of his model under both random effects and fixed effects assumptions. In contrast, we discuss partial identification of a model that is not additively separable in U_{it} under fixed effects assumption.

The discussion in the following is technical in nature. For this discussion, we assume $T \geq 3$ and focus on the case $T = 3$. Let $\mathcal{X} = \text{Supp}(X_{it})$. For simplicity, we assume that this support is the same for each t. Let h_* denote the inverse of m_*. Then

$$X_{it}h_*(Y_{it-1}) - X_{it-1}h_*(Y_{it}) = (X_{it} - X_{it-1})\alpha_* + X_{it}U_{*it-1} - X_{it-1}U_{*it}. \tag{51}$$

Let $X_i = (X_{i1}^\intercal, X_{i2}^\intercal, X_{i3}^\intercal)^\intercal$, $Y_i = (Y_{i1}^\intercal, Y_{i2}^\intercal, Y_{i3}^\intercal)^\intercal$. Let $X_{i(-2)} = (X_{i1}^\intercal, X_{i3}^\intercal)^\intercal$, $Y_{i(-2)} = (X_{i1}^\intercal, X_{i3}^\intercal)^\intercal$, and $X_{i(-t)}, Y_{i(-t)}$ for $t = 1, 3$ be defined analogously. We maintain the following assumptions.

Assumption 5 *(i)* $(Y_i, X_i, \alpha_{*i})_{i=1}^n$ *is a random sample.*

(ii) *For* $t = 2, 3$ *and* $x \neq 0$, $\text{Supp}(Y_{it}|X_{it} = X_{it-1} = x) = \text{Supp}(Y_{it-1}|X_{it} = X_{it-1} = x)$, *and the joint density* $f_{X_{it-1}, X_{it}}(x, x) > 0 \ \forall x \in \mathcal{X}$.

(iii) *(Location normalization)* $h_*(0) = 0$.

(iv) *(Fixed effect matters for almost all treatment values) For some* $t \in \{2, 3\}$, *say* $t = 2$, $\mathbb{P}(X_t\beta_{2*} + 1 = 0) = 0$. *(Below, we assume that this is the case for* $t = 2$*).*

(v) *For* $t = 1, 2, 3$, $f_{U_{*it}|X_{it}, \alpha_{*i}, X_{i(-t)}, U_{*i(-t)}}(u_t|x, a, x_{(-t)}, u_{(-t)}) = f_{U_{*it}|X_{it}}(u_t|x)$ *a.s.*.

(vi) *For $t = 1, 2, 3$, $\mathbb{E}(U_{*it}|X_{it}) = 0$.*

(vii) *For each $s \in \mathbb{R}$, and each t, with probability 1, conditional characteristic function of $U_{*it}|X_{it}$ evaluated at s is different from 0.*

(viii) *For $t = 1, 2, 3$, $\mathbb{E}[|h_*(Y_{it})||X_i = x_i]$ and $\mathbb{E}[|U_{*it}||X_i = x_i]$ are bounded for all t and x_i.*

(ix) *Conditional distribution of α_{*i} given X_i is continuous.*

(x) *The density functions $f_{U_{*it}|X_{it}}(u|x)$, $f_{\alpha_{*i}|X_i}(a|x_i)$ are continuous in the continuously distributed components of x and x_i for all $a \in \mathbb{R}$ and $u \in \mathbb{R}$.*

(xi) *There exists $r > 0$ such that $B(0, r) \in Supp(Y_{t-1}, Y_t)$, where $B(0, r)$ denotes a ball of radius r.*

Parts (v)-(x) of this assumption are versions of Assumption ID of Evdokimov (2010) adapted to our setting. Part (ii) is a support condition. Part (iii) is a location normalization. Part (iv) of this assumption is used for derivation of the distribution of $\alpha_*|X_t$. Without this assumption, we could calculate the distribution of $\alpha_{i*}|X_{it} = x_t, X_{it-1} = x_{t-1}$ for $x_t \neq x_{t-1}$.In the analysis below, the subscript i is dropped.

Below, we present a constructive algorithm of finding the conditional characteristic functions $\phi_{U_{*1}|X_1}$, $\phi_{U_{*2}|X_2}$, $\phi_{U_{*3}|X_3}$, $\phi^m_{\alpha_m|X_1,X_2,X_3}$ of $(U_{*1}, U_{*2}, U_{*3}, \alpha)$ as well as $\beta_{1*} = (\beta_{1*}, \beta_{2*})$ if we are given the knowledge of the true m_*. In other words, we present a mapping $\theta_m := \left(\phi^m_{U_{m1}|X_1}, \phi^m_{U_{m2}|X_2}, \phi^m_{U_{m3}|X_3}, \phi^m_{\alpha_m|X_1,X_2,X_3}, \beta_m\right)$ such that θ_m coincides with the truth if $m = m_*$. This implies that if $m = m_*$, we can identify the conditional distribution $F_{Y^m|X}$ of $Y^m = (Y_1^m, Y_2^m, Y_3^m)^\top$ given $X = (X_1, X_2, X_3)^\top$, where $Y_t^m := m(\alpha_m + X_t\beta_{1m} + X_t\beta_{2m}\alpha_n + U_{tm})$. Our algorithm can be understood as consisting of the results presented in Lemmas 1 - 5 below.

Lemma 1 *Under Assumptions (1) and (5),*

$$H_{*2}(s, x) = \exp\left(\begin{array}{c} \int_0^s \frac{\mathbb{E}[i\{h_*(Y_2)-h_*(Y_1)\}\exp(it\{h_*(Y_2)-h_*(Y_3)\})|X_1=X_2=X_3=x]}{\mathbb{E}[\exp(it\{h_*(Y_2)-h_*(Y_3)\})|X_1=X_2=X_3=x]}dt \\ +is\mathbb{E}[h_*(Y_1) - h_*(Y_2)|X_1 = X_2 = X_3 = x] \end{array} \right),$$

where

$$H_{*t}(s, x) := \mathbb{E}\left[\exp(isU_{*t})|X_1 = X_2 = X_3 = x\right] = \mathbb{E}\left[\exp(isU_{*t})|X_t = x\right].$$

Proof. We use the same arguments as in the Proof of Lemma 1 of Evdokimov (2010). For this argument we rely on the previous step and the equivalence between characteristic functions and distributions of random variables/vectors.

Let $x \neq 0$, and

$$\begin{aligned} G_*(s_1, s_2, x) &:= \phi_{h_*(Y_2)-h_*(Y_1), h_*(Y_3)-h_*(Y_1)|X_1=X_2=X_3=x}(s_1, s_2) \\ &= \mathbb{E}\left[\exp\{is_1(U_{*2} - U_{*1}) + is_2(U_{*3} - U_{*1})\}|X_1 = X_2 = X_3 = x\right] \\ &= \mathbb{E}\left[\exp\left(-i\left(s_1 + s_2\right)U_{*1} + is_1U_{*2} + is_2U_{*3}\right)|X_1 = X_2 = X_3 = x\right] \end{aligned}$$

Under Assumption (v) above, we can repeatedly obtain

$$\begin{aligned} &\mathbb{E}\left[\exp\left(-i\left(s_1 + s_2\right)U_{*1} + is_1U_{*2} + is_2U_{*3}\right)|\, X_1 = X_2 = X_3 = x, \alpha_*, U_{*1}, U_{*2}\right] \\ &= \exp\left(-i\left(s_1 + s_2\right)U_{*1} + is_1U_{*2}\right)\mathbb{E}\left[\exp\left(is_2U_{*3}\right)|\, X_1 = X_2 = X_3 = x, \alpha_*, U_{*1}, U_{*2}\right] \\ &= \exp\left(-i\left(s_1 + s_2\right)U_{*1} + is_1U_{*2}\right)\mathbb{E}\left[\exp\left(is_2U_{*3}\right)|\, X_3 = x\right], \end{aligned}$$

$$\begin{aligned} &\mathbb{E}\left[\exp\left(-i\left(s_1 + s_2\right)U_{*1} + is_1U_{*2} + is_2U_{*3}\right)|\, X_1 = X_2 = X_3 = x, \alpha_*, U_{*1}\right] \\ &= \exp\left(-i\left(s_1 + s_2\right)U_{*1}\right)\mathbb{E}\left[\exp\left(is_2U_{*3}\right)|\, X_3 = x\right]\mathbb{E}\left[\exp\left(is_1U_{*2}\right)|\, X_1 = X_2 = X_3 = x, \alpha_*, U_{*1}\right] \\ &= \exp\left(-i\left(s_1 + s_2\right)U_{*1}\right)\mathbb{E}\left[\exp\left(is_1U_{*2}\right)|\, X_2 = x\right]\mathbb{E}\left[\exp\left(is_2U_{*3}\right)|\, X_3 = x\right], \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}\left[\exp\left(-i\left(s_1 + s_2\right)U_{*1} + is_1U_{*2} + is_2U_{*3}\right)|\, X_1 = X_2 = X_3 = x, \alpha_*\right] \\ &= \mathbb{E}\left[\exp\left(-i\left(s_1 + s_2\right)U_{*1}\right)|\, X_1 = x\right]\mathbb{E}\left[\exp\left(is_1U_{*2}\right)|\, X_2 = x\right]\mathbb{E}\left[\exp\left(is_2U_{*3}\right)|\, X_3 = x\right], \end{aligned} \tag{52}$$

which yields

$$G_*(s_1, s_2, x) = H_{*1}(-s_1 - s_2, x)\, H_{*2}(s_1, x)\, H_{*3}(s_2, x),$$

where

$$H_{*t}(s, x) := \mathbb{E}\left[\exp\left(isU_{*t}\right)|X_1 = X_2 = X_3 = x\right] = \mathbb{E}\left[\exp\left(isU_{*t}\right)|X_t = x\right].$$

(Given $X_1 = X_2 = X_3 = x$, G_* and H_{*t} are deterministic functions of h_*.) Therefore,

$$\frac{\frac{\partial G_*(s_1, s_2, x)}{\partial s_1}}{G_*(s_1, s_2, x)} = -\frac{\frac{\partial H_{*1}(-s_1 - s_2, x)}{\partial s_1}}{H_{*1}(-s_1 - s_2, x)} + \frac{\frac{\partial H_{*2}(s_1, x)}{\partial s_1}}{H_{*2}(s_1, x)},$$

from which we obtain

$$\frac{\frac{\partial G_*(s, -s, x)}{\partial s_1}}{G_*(s, -s, x)} = -\frac{\partial H_{*1}(0, x)/\partial s}{H_{*1}(0, x)} + \frac{\partial H_{*2}(s, x)/\partial s}{H_{*2}(s, x)}, \tag{53}$$

Note that the first term on the right of (53) can be understood to be $-i\mathbb{E}[U_{*1}|\, X_1 = X_2 = X_3 =$

$x]$. Using the assumption that $\mathbb{E}[U_{*2}|X_2 = x] = 0$, equation (53) can be rewritten

$$\frac{\frac{\partial G_*(s,-s,x)}{\partial s_1}}{G_*(s,-s,x)} = -i\mathbb{E}[h_*(Y_1) - h_*(Y_2)|X_1 = X_2 = X_3 = x] + \frac{\partial H_{*2}(s,x)/\partial s}{H_{*2}(s,x)}.$$

Therefore, we can write

$$\begin{aligned} H_{*2}(s,x) &= \exp\left(\int_0^s \frac{\partial G_*(t,-t,x)/\partial s_1}{G_*(t,-t,x)}dt + is\mathbb{E}[h_*(Y_1) - h_*(Y_2)|X_1 = X_2 = X_3 = x]\right) \\ &= \exp\left(\begin{array}{c} \int_0^s \frac{\mathbb{E}[i\{h_*(Y_2)-h_*(Y_1)\}\exp(it\{h_*(Y_2)-h_*(Y_3)\})|X_1=X_2=X_3=x]}{\mathbb{E}[\exp(it\{h_*(Y_2)-h_*(Y_3)\})|X_1=X_2=X_3=x]}dt \\ +is\mathbb{E}[h_*(Y_1) - h_*(Y_2)|X_1 = X_2 = X_3 = x] \end{array}\right), \end{aligned}$$

where the second equality uses the alternative characterization of $G_*(s,-s,x)$ and $\partial G_*(s,-s,x)/\partial s_1$ below:

$$\begin{aligned} G_*(s,-s,x) &= \mathbb{E}\left[\exp\left(is\{h_*(Y_2) - h_*(Y_3)\}\right)|X_1 = X_2 = X_3 = x\right], \\ \frac{\partial G_*(s,-s,x)}{\partial s_1} &= \mathbb{E}\left[i\{h_*(Y_2) - h_*(Y_1)\}\exp\left(is\{h_*(Y_2) - h_*(Y_3)\}\right)|X_1 = X_2 = X_3 = x\right]. \end{aligned}$$

■

Lemma 2 *Under Assumptions (1) and (5),*

$$\mathbb{E}\left[\exp\left(isU_{*1}\right)|X_1 = x\right] = \frac{\mathbb{E}\left[\exp\{is(h_*(Y_1) - h_*(Y_2))\}|X_1 = X_2 = x\right]}{\mathbb{E}\left[\exp\left(-isU_{*2}\right)|X_2 = x\right]}.$$

Proof. Note that under the assumption that $U_{*1} \perp\!\!\!\perp U_{*2}|X_1 = x, X_2 = x, X_3 = x$ and using equation (51), we have

$$\mathbb{E}\left[\exp\left(isU_{*1}\right)|X_1 = x\right] = \frac{\mathbb{E}\left[\exp\{is(h_*(Y_1) - h_*(Y_2))\}|X_1 = X_2 = x\right]}{\mathbb{E}\left[\exp\left(-isU_{*2}\right)|X_2 = x\right]}.$$

■

Lemma 3 *Under Assumptions (1) and (5),*

$$\begin{aligned} &\mathbb{E}\left[\alpha_*|\, X_1 = x_1, X_2 = x_2, X_3 = x_3\right] \\ = \;& \frac{1}{i(x_2 - x_1)}\,\frac{\partial}{\partial s}\left(\frac{\mathbb{E}\left[\exp\left\{is\left(X_2h_*(Y_1) - X_1h_*(Y_2)\right)\right\}|\, X_1 = x_1, X_2 = x_2, X_3 = x_3\right]}{\mathbb{E}\left[\exp\{isx_2U_{*1}\}|\, X_1 = x_1\right]\mathbb{E}\left[\exp\{-isx_1U_{*2}\}|\, X_2 = x_2\right]}\right)\Bigg|_{s=0} \end{aligned}$$

Proof. For $x_2 \neq x_1$, using equation (51) and the previous step, we have

$$\frac{\mathbb{E}\left[\exp\left\{is\left(X_2h_*(Y_1) - X_1h_*(Y_2)\right)\right\}|\, X_2 = x_2, X_1 = x_1\right]}{\mathbb{E}\left[\exp\{isx_2U_{*1}\}|\, X_1 = x_1\right]\mathbb{E}\left[\exp\{-isx_1U_{*2}\}|\, X_2 = x_2\right]} = \mathbb{E}\left[\exp\{is(x_2 - x_1)\alpha_*|\, X_2 = x_2, X_1 = x_1\right].$$

Differentiating both sides with respect to s and then evaluating that derivative at $s = 0$ yields

$(x_2 - x_1)\mathbb{E}[\alpha_*|X_2 = x_2, X_1 = x_1]$ or equivalently, $\mathbb{E}[\alpha_*|X_2 = x_2, X_1 = x_1]$. In this step, we could alternatively condition on $X_1 = x_1, X_2 = x_2, X_3 = x_3$ with at least two of x_1, x_2, x_3 being different from each other. Using the same arguments, we could obtain $\mathbb{E}[\alpha_*|X_1 = x_1, X_2 = x_2, X_3 = x_3]$.

■

Lemma 4 *Under Assumptions (1) and (5), we can identify β_* by solving $B_*\beta_* = A_*$, where*

$$A_* = \begin{pmatrix} \frac{\mathbb{E}(h_*(Y_2) - \alpha_*|X_1 = x_1, X_2 = x_2, X_3 = x_3)}{x_2} \\ \frac{\mathbb{E}(h_*(Y_2) - \alpha_*|X_1 = \tilde{x}_1, X_2 = \tilde{x}_2, X_3 = \tilde{x}_3)}{\tilde{x}_2} \end{pmatrix},$$
$$B_* = \begin{bmatrix} 1 & \mathbb{E}(\alpha_*|X_1 = x_1, X_2 = x_2, X_3 = x_3) \\ 1 & \mathbb{E}(\alpha_*|X_1 = \tilde{x}_1, X_2 = \tilde{x}_2, X_3 = \tilde{x}_3) \end{bmatrix},$$
$$\beta_* = \begin{pmatrix} \beta_{1*} \\ \beta_{2*} \end{pmatrix}.$$

Remark 16 *We have $B_*\beta_* = A_*$, with B_* invertible under our assumptions. Since the matrices $A_*, B*$ are deterministic functions of h_*, so is β_*.*

Lemma 5 *Under Assumptions (1) and (5), we have*

$$\mathbb{E}\left[\exp\left(is\{1 + X_2\beta_{2*}\}\alpha_*\right)|\, X_1, X_2, X_3\right] = \frac{\mathbb{E}[\exp\left(is\{h_*(Y_2) - X_2\beta_{1*}\}\right)|\, X_1, X_2, X_3]}{\mathbb{E}[\exp\left(isU_{*2}\right)|X_2]}.$$

Proof. It follows from

$$\mathbb{E}[\exp\left(is\{h_*(Y_2) - X_2\beta_{1*}\}\right)|X_1, X_2, X_3] = \mathbb{E}[\exp\left(is\{1 + X_2\beta_{2*}\}\alpha_*\right)|X_1, X_2, X_3]\mathbb{E}[\exp\left(isU_{*2}\right)|X_2].$$

■

Remark 17 *By Assumption 5(iv), by varying s and using the inversion formula, we obtain the distribution of $\alpha_*|\,(X_1, X_2, X_3)$, which is also a deterministic function of h_*.*

When $X_t = X_{t-1} = x$ with $x \neq 0$, equation (51) becomes

$$h_*(Y_{t-1}) - h_*(Y_t) = [U_{*t-1} - U_{*t}]. \tag{54}$$

Suppose there is another model $(m, \beta_m, \alpha_m, \{U_{mt}\}_{t=1}^T)$ that satisfies equation (50) and Assumption 5 such that $U_{mt-1} - U_{mt} = U_{*t-1} - U_{*t}$ *a.s.*, that is $h_*(Y_{t-1}) - h_*(Y_t) = h_m(Y_{t-1}) - h_m(Y_t)$ *a.s.*. Letting y_t go to y_{t-1} (or vice versa) conditional on the event $X_t = X_{t-1} = x$ with $x \neq 0$, then we have $h'_*(y_{t-1}) = h'_m(y_{t-1})$. Note that by Assumption 5 and the fact that Y_t is continuously distributed, we could do this for each realization of Y_{t-1}. then h_* and h_m would have to be the same up to a location normalization. If we started with another m function, these steps would

give us

$$\theta_m := \left(\phi^m_{U_{m1}|X_1}, \phi^m_{U_{m2}|X_2}, \phi^m_{U_{m3}|X_3}, \phi^m_{\alpha_m|X_1,X_2,X_3}, \beta_m\right).$$

Based on θ_m thus identified, we can identify the conditional distribution $F_{Y^m|X}$ of $Y^m = (Y_1^m, Y_2^m, Y_3^m)^\top$ given $X = (X_1, X_2, X_3)^\top$, where $Y_t^m = m\left(\alpha_m + X_t\beta_{1m} + X_t\beta_{2m}\alpha_n + U_{tm}\right)$. The partially identified set of m is then given by $\mathcal{M} := \{m \in \mathcal{C} : F_{Y^m|X} = F_{Y|X}\ a.s.\}$ where $\mathcal{C}$ is the set of strictly increasing and continuously differentiable functions m.

The specification $h(\cdot,\cdot)$ is falsified if there is no $h^{-1}(\cdot,\delta)$ that belongs to the identified set. These arguments lead us to the following results:

Proposition 13 *Suppose assumptions (1) and (5), $m_* \in \mathcal{M}$.*

A.9 A Model of Endogenous Teacher-Student Match Effects

The following shows that a model of endogenous teacher-student match effects generates a production function for student skills that depends on a scalar teacher-specific unobservable that enters linearly and interacts with a student's exogenous characteristics, as in equation (24). In the model, teachers are heterogeneous in productivity, students differ in their characteristics, and a classroom teacher chooses a teaching effort level for each student.

Let θ_j represent teacher j's idiosyncratic productivity and x_s summarize the characteristics of student s. For ease of exposition, x_s can represent i's lagged normalized score, which is continuously distributed. Assuming skill production is linear in effort and a quadratic effort cost, teacher j's optimal effort choice for student s is given by:

$$e_j^*(x_s) = \theta_j \frac{w(x_s)}{\kappa(x_s)}$$

where $w(x_s)$ is the weight student i receives in the teacher's objective function and $\kappa(x_s)$ shifts the marginal cost of educating student s. We assume that the weight and cost functions are differentiable. Thus, effort is increasing in the weight assigned a student; is decreasing in the cost of educating a student; and a higher productivity teacher produces the same output at lower cost, all else equal.

We now consider a Taylor approximation of the optimal effort choice around $x_s = 0$ to derive an equation that is linear in x_s:

$$\begin{aligned} e_j^*(x_s) &\approx \theta_j \left[\frac{w(0)}{\kappa(0)} + \left(\frac{w'(0,0)}{\kappa(0)} - \frac{w(0)\kappa'(0)}{\kappa(0)^2}\right) x_s\right] \\ &= \theta_j \left[\frac{w(0)}{\kappa(0)} + \left(\frac{w'(0)}{w(0)} - \frac{\kappa'(0)}{\kappa(0)}\right) \frac{w(0)}{\kappa(0)} x_s\right] \\ &= \alpha_j + \underbrace{\left[\frac{w'(0)}{w(0)} - \frac{\kappa'(0)}{\kappa(0)}\right]}_{\beta} \alpha_j x_s \end{aligned} \tag{55}$$

where $\alpha_j = \theta_j \frac{w(0)}{\kappa(0)}$. This yields an estimating equation for skills analogous with equation (24):

$$y_{js} = \alpha_j + \beta \alpha_j x_s + \epsilon_{js}$$

where ϵ_s contains all other determinants of skills.[20] Thus, a simple model of endogenous teacher-student match effects naturally generates the scalar teacher fixed effect model considered in our empirical application.

Moreover, the parameters of the estimating equation have intuitive interpretations in terms of the model's primitives. Teacher quality α_j, for example, is idiosyncratic across teachers, but also reflects effort supplied in educating an average student. β depends on the (normalized) slopes of the weight and cost functions and may be positive or negative. As such, β, which governs the sign and magnitude of the match effects, may depend on workplace features of education production, such as accountability policy. If a student's weight increases in characteristic x_s or their cost decreases in x_s, teacher quality will endogenously be more effective (all else equal). Further, it is easy to see that the homogeneous—no match effects—teacher value-added model follows from the special case where both the weights and costs do not depend on student characteristics. In other words, if weights or effort costs differ across students, we should expect to find match effects in the data *even when teacher quality can be summarized by a scalar index.*

A.10 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

[20] This likely would include x_s; this framework here abstracts from household inputs to education.

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.11 Estimating Equation

The full estimating equation in Section 7 when interactions with lagged score, economic disadvantage, and underrepresented minority status are simultaneously included is given by:

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

where z^L_{st} is student s's lagged score, z^E_{st} is an indicator for economic disadvantage, and z^U_{st} is an indicator for underrepresented minority; $\bar{z}_{jt}$ are the respective classroom averages. The π parameters are reduced-form (non-linear) combinations of β and γ, but their relationship is not imposed in estimation.

A.12 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: Teacher Quality by Quintiles of Classroom Disadvantage

Class average			Teacher quality (math)			Teacher quality (reading)		
Lag score (math)	Econ. dis.	URM	Het.	No het.	Diff.	Het.	No het.	Diff.
0.41	0.12	0.18	0.06	0.05	0.006	0.05	0.05	-0.003
0.20	0.32	0.23	0.02	0.02	0.002	0.02	0.02	-0.005
0.04	0.46	0.30	-0.01	-0.01	-0.003	-0.01	0.00	-0.009
-0.13	0.61	0.42	-0.03	-0.02	-0.006	-0.03	-0.02	-0.012
-0.39	0.83	0.68	-0.04	-0.04	0.000	-0.02	-0.05	0.030

Table A3: 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)