Review of Economic Studies (2018) **85**, 1824–1851 doi:10.1093/restud/rdx052
© The Author 2017. Published by Oxford University Press on behalf of The Review of Economic Studies Limited.
Advance access publication 6 September 2017

Non-parametric Panel Data Models with Interactive Fixed Effects

JOACHIM FREYBERGER
University of Wisconsin - Madison

First version received June 2014*; Editorial decision June* 2017*; Accepted September* 2017 (*Eds.*)

This article studies non-parametric panel data models with multidimensional, unobserved individual effects when the number of time periods is fixed. I focus on models where the unobservables have a factor structure and enter an unknown structural function non-additively. The setup allows the individual effects to impact outcomes differently in different time periods and it allows for heterogeneous marginal effects. I provide sufficient conditions for point identification of all parameters of the model. Furthermore, I present a non-parametric sieve maximum likelihood estimator as well as flexible semiparametric and parametric estimators. Monte Carlo experiments demonstrate that the estimators perform well in finite samples. Finally, in an empirical application, I use these estimators to investigate the relationship between teaching practice and student achievement. The results differ considerably from those obtained with commonly used panel data methods.

Key words: Panel data, Multidimensional individual effects, Factor model, Non-parametric identification.

JEL Codes: C14, C23

1. INTRODUCTION

A standard linear fixed effects panel data model allows for a scalar unobserved individual effect, which may be correlated with explanatory variables. Consequently, by making use of panel data, a researcher may allow for endogeneity without the need for an instrumental variable. However, a scalar unobserved individual effect, which enters additively, imposes two important restrictions. To illustrate these restrictions, suppose that the observed outcome Y_{it} denotes the test score of student i in test t. Here the researcher could either observe the same student taking tests in different time periods or, as in many empirical applications, the researcher could observe several subject specific tests for the same student.[1] In these applications the individual effect typically represents unobserved ability of student i and the explanatory variables include student and teacher characteristics. Since the individual effect is a scalar and constant across t, the first main restriction is that if one student has a higher individual effect than another student with the same observed characteristics, then the student with the higher individual effect also has a higher

1. For example, using subject specific tests, Dee (2007) analyses whether assignment to a same-gender teacher has an influence on student achievement. Clotfelter *et al*. (2010) and Lavy (2016) investigate the relationship between teacher credentials and student achievement and teaching practice and student achievement, respectively.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

expected test outcome in *all* tests. Hence, it is not possible that student i has abilities such that she is better in mathematics, while student j is better in English. The second main restriction is that the model does not allow for interactions between individual effects and explanatory variables. Therefore, in the previous example, the linear fixed effects model implicitly assumes that the effect of a teacher characteristic on test scores does not depend on students' abilities.

To allow for these empirically relevant features, in this article I study panel data models with multidimensional individual effects and marginal effects that may depend on these individual effects. Specifically, I consider models based on

$$Y_{it} = g_t\left(X_{it}, \lambda_i' F_t + U_{it}\right), \qquad i = 1, \ldots, n, t = 1, \ldots, T, \tag{1}$$

where Y_{it} is a scalar outcome variable, g_t is an unknown structural function, $X_{it} \in \mathbb{R}^{d_x}$ is a vector of explanatory variables, $\lambda_i \in \mathbb{R}^R$ and $F_t \in \mathbb{R}^R$ are unobserved vectors, and U_{it} is an unobserved random variable. The explanatory variables $X_i = (X_{i1}, \ldots, X_{iT})$ may be continuous or discrete and X_i may depend on the individual effects λ_i. In the previous example, λ_i accounts for different dimensions of unobserved abilities of student i and F_t is the importance of the abilities for test t. Hence, both the returns to the various abilities and the relative importance of each ability on the outcome can change across tests. Thus, some students may have higher expected outcomes in mathematics, while others may have higher expected outcomes in English, without changes in covariates. Furthermore, since the structural functions are unknown, the model allows for a flexible relationship between Y_{it} and X_{it}, and the effect of X_{it} on Y_{it} may depend on λ_i. A semiparametric special case of the model, which is covered by the results in this paper, is $\alpha_t(Y_{it}) = X_{it}'\beta_t + \lambda_i' F_t + U_{it}$ where α_t is an unknown strictly increasing transformation of Y_{it}. Such a model is particularly appealing when Y_{it} are test scores, because test scores do not have a natural metric and any increasing transformation of them preserves the same ranking of students (see Cunha and Heckman, 2008). Thus, next to estimating the slope coefficients, a researcher can allow for an unknown transformation of the test scores. Other special cases of (1) include a linear factor model, where g_t is linear, as well as the standard linear fixed effects model with both scalar individual effects and time dummies. Notice that while a linear factor model allows for multiple individual effects, it does not allow for heterogeneous marginal effects.

The models studied in this article are appealing in a variety of empirical applications where unobserved heterogeneity is not believed to be one dimensional and time homogeneous, and a researcher wants to allow for a flexible relationship between Y_{it}, X_{it}, and the unobservables. Examples include estimating the returns to education or the effect of union membership on wages (where λ_i represents different unobserved abilities and F_t their price at time t), estimating production functions (where λ_i can capture different unobserved firm specific effects), and cross country regressions (where F_t denotes common shocks and λ_i the heterogeneous impacts on country i).[2]

This article presents sufficient conditions for point identification of all parameters of models based on outcome Equation (1) when T is fixed and the number of cross-sectional units is large. In the previous example, where T represents the number of tests, I therefore only require a small number of tests for each student. The identified parameters include the structural functions g_t, the number of individual effects R, the vectors F_t, and the distribution of the individual effects conditional on the covariates.[3] Identification of these parameters immediately

2. For more examples of factor models in economics see Bai (2009) and references therein.

3. The factor structure of the unobservables is commonly called interactive fixed effects due to the interaction of λ_i and F_t. The vector F_t is usually referred to as the factors, while λ_i is called the loadings. I use this terminology because I do not impose a parametric assumption on the dependence between λ_i and X_{it}. Graham and Powell (2012) provide a discussion on the difference between *fixed effects* and *correlated random effects*.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

implies identification of economically interesting features such as average and quantile structural functions. Although T is fixed, I require that $T \geq 2R+1$ so that for a given T only models with at most $(T-1)/2$ factors are point identified, which is also a standard condition in linear factor models. The main result in the article is for continuously distributed outcomes, where my assumptions are natural extensions of those in a linear factor model, but the identification arguments are substantially different. As in the linear model, the assumptions rule out lagged dependent variables as regressors. However, I discuss extensions to allow for lagged dependent variables as regressors, as well as discretely distributed outcomes, in the Supplementary Appendix (see Remark 3).

I then show that a non-parametric sieve maximum likelihood estimator estimates all parameters consistently. Since the estimator requires estimating objects which might be high dimensional in applications, such as the density of $\lambda_i \,|\, X_i$, this paper also provides a flexible semiparametric estimator, where I reduce the dimensionality of the estimation problem by assuming a location and scale model for the conditional distributions. I provide conditions under which the finite dimensional parameters are $\sqrt{n}$ consistent and asymptotically normally distributed, and I also describe an easy to implement fully parametric estimator.

In an empirical application, I study the relationship between teaching practice and student achievement, where Y_{it} are different mathematics and science test scores for each student i. The main regressors are measures of traditional and modern teaching practice for each class a student attends, constructed from a student questionnaire. Traditional and modern teaching practices are associated with lectures/memorizing and group work/explanations, respectively. I estimate marginal effects of teaching practice, on mathematics and science test scores, for different levels of students' abilities and find that the semiparametric two factor model yields substantially different conclusions than a linear fixed effects model.

Many recent papers in the non-parametric panel data literature are related to the models I consider. First, several papers make use of some form of time homogeneity to achieve identification, do not restrict the dependence of U_{it} over t or the distribution of $\lambda_i \,|\, X_i$, and achieve identification of average or quantile effects. Papers in this category include Graham and Powell (2012), Hoderlein and White (2012) and Chernozhukov *et al*. (2013).[4] Chamberlain (1992) analyses common parameters in random coefficient models. Arellano and Bonhomme (2012) extend his analysis and restrict the dependence of U_{it} over t to obtain identification of the variance and the distribution of the random coefficients. While all of these papers allow for multiple individual effects and heterogeneous marginal effects, the time homogeneity assumptions imply that the ranking of individuals based on $E[Y_{it} \,|\, X_i, \lambda_i]$ cannot change over t without a change in X_{it}. Contrarily, compared to those papers, (1) makes stronger assumptions on the dimension of λ_i, assumes that λ_i affects Y_{it} through an index, and requires independence of U_{it} over t for identification (see Section 2.2 for more details). It therefore rules out random coefficients for example. Thus, (1) is most useful if one believes that λ_i has a different effect on Y_{it} for different t and is willing to put some structure on the unobservables. Bester and Hansen (2009) do not impose time homogeneity and instead restrict the distribution of $\lambda_i \,|\, X_i$. Altonji and Matzkin (2005) require an external variable, which they construct in panels by restricting the distribution of $\lambda_i \,|\, X_i$. Wilhelm (2015) analyses a non-parametric panel data model with measurement error and an additive scalar individual effect. Evdokimov (2010, 2011) assumes that U_{it} is independent over t and he uses identification arguments that are related to those in the measurement error

4. Scalar additive or multiplicative time effects are allowed in some of these papers.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

literature. He provides identification results in non-separable models with a scalar heterogeneity term as well as a novel conditional deconvolution estimator.[5]

I also make use of measurement error type arguments instead of relying on time homogeneity or restricting the distribution of $\lambda_i | X_i$. Specifically, I build on the work of Hu (2008), Hu and Schennach (2008) and Cunha *et al.* (2010). Hu and Schennach (2008) study a non-parametric measurement error model with instruments. The connection to (1) is that λ_i can be seen as unobserved regressors, a subset of the outcomes represents observed and mismeasured regressors, and another subset of outcomes serves as instruments. Cunha *et al.* (2010) apply results in Hu and Schennach (2008) to a measurement model of the general form $Y_{it} = g_t(\lambda_i, U_{it})$. Compared to the general model, I use a more restrictive outcome equation to reduce the dimensionality of the estimation problem, which may be appealing in empirical applications. As a consequence, two main identifying assumptions in Cunha *et al.* (2010) cannot be used in my setting, which changes important arguments in the identification proofs. In particular, one of their main identifying assumption fixes a measure of location of the distribution of a subset of outcomes given λ_i.[6] In my model, such an assumption would impose very strong restrictions on g_t. Instead, I use the relation between Y_{it} and λ_i delivered by (1), combined with arguments from linear factor models and single index models. Moreover, Cunha *et al.* (2010) impose an assumption on the conditional distribution of the outcomes, which does not hold with my factor structure and $T = 2R+1$.[7] I instead show that interchangeability of outcomes can be used to obtain identification with $T = 2R+1$. These results require stronger independence assumptions compared to Cunha *et al.* (2010), but some of these assumptions also serve as sufficient conditions for their completeness assumptions and are used to identify average and quantile structural functions. Finally, I consider extensions to allow for an unknown R and lagged dependent variables as regressors.

This article is also related to a vast literature on linear factor models, which are well understood and can deal with multiple unobserved individual effects.[8] Non-linear models of the form $Y_{it} = g(X_{it}) + \lambda_i' F_t + U_{it}$ have been studied by Su and Jin (2012) and Huang (2013) when $n, T \to \infty$. A drawback of additively separable models is that they impose homogeneous marginal effects. The analysis in these papers is tailored to additively separable models. For example, estimation in Bai (2009) is based on the method of principal components.

The remainder of the article is organized as follows. Section 2 outlines the identification arguments. Section 3 discusses different ways to estimate the model. Sections 4 and 5 contain the empirical application and Monte Carlo simulation results, respectively. Finally, Section 6 concludes. The proofs of the main results are in the Appendix. Additional material is in a Supplementary Appendix with Section numbers S.1, S.2, etc..

Notation: To simplify the notation, I drop the subscript i from all random variables in the remainder of the article and write the outcome equation as $Y_t = g_t\left(X_t, \lambda' F_t + U_t\right)$. For each t, let

5. Many of these papers also assume some form of strict exogeneity for their main results (Assumption N4). Exceptions are Altonji and Matzkin (2005), who instead assume an exchangeability condition, and Chernozhukov *et al.* (2013).

6. See their Assumption (v) of Theorem 2. Hu and Schennach (2008) fix a measure of location of the distribution of the measurement error.

7. See their Assumption (iv) of Theorem 2. The assumption holds with a factor structure when $T \geq 3R$ and can also hold with $T = 2R+1$ if the unobservables do not have a factor structure.

8. The theoretical literature on linear factor models includes Heckman and Scheinkman (1987), Holtz-Eakin *et al.* (1988), Ahn *et al.* (2001), Bai and Ng (2002), Bai (2003), Andrews (2005), Pesaran (2006), Bonhomme and Robin (2008), Bai (2009), Ahn *et al.* (2013), Bai (2013), and Moon and Weidner (2015). Factor models have also been used in applications related to the one in this paper, including Carneiro *et al.* (2003), Heckman *et al.* (2006), Cunha and Heckman (2008), Cunha *et al.* (2010), and Williams *et al.* (2010).

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

$\mathcal{X}_t \subseteq \mathbb{R}^K$ and $\mathcal{Y}_t \subseteq \mathbb{R}$ be the supports of X_t and Y_t, respectively. Let $\Lambda \subseteq \mathbb{R}^R$ be the support of λ. Define $X = (X_1, \ldots, X_T)$ and define Y and U analogously. Let $\mathcal{X}$ and $\mathcal{Y}$ be the supports of X and Y, respectively. The conditional pdf of any random variable $W \mid V$ is denoted by $f_{W|V}(w; v)$ and the marginal pdf by $f_W(w)$.

2. IDENTIFICATION

In this section, I assume that R is known. I consider identification of the number of factors in Section S.1.1 of the Supplementary Appendix. Before discussing the general model, I provide intuition for the main result by showing identification of a linear model, where the main arguments go back to Madansky (1964) and are very similar to those of Heckman and Scheinkman (1987).

2.1. *Preliminaries: linear factor models*

I consider a linear factor model with $T = 5$ and $R = 2$, where X_t is a scalar and

$$Y_t = X_t \beta_t + \lambda_1 F_{t1} + \lambda_2 F_{t2} + U_t. \tag{2}$$

I make the following assumptions.

Assumption L1. *$F_4 = \begin{pmatrix} 1 & 0 \end{pmatrix}'$ and $F_5 = \begin{pmatrix} 0 & 1 \end{pmatrix}'$.*

Assumption L2. *$E[U_t \mid X, \lambda] = 0$ for all $t = 1, \ldots 5$.*

Assumption L3. *$U_1, \ldots, U_5, \lambda$ are uncorrelated conditional on X.*

Assumption L4. *The 2×2 matrix $\begin{pmatrix} F_t & F_s \end{pmatrix}$ has full rank for all $s \neq t$.*

Assumption L5. *The 2×2 covariance matrix of λ has full rank conditional on X.*

Assumption L6. *For any $t_1 \in \{1, \ldots, 5\}$ there exists $t_2, t_3 \in \{1, \ldots, 5\} \setminus t_1$ such that $Var(X_{t_1} \mid X_{t_2}, X_{t_3}) > 0$.*

Assumption L1 is a normalization needed because for any $R \times R$ invertible matrix H it holds that $\lambda' F_t = \lambda' H H^{-1} F_t = (H'\lambda)'(H^{-1} F_t) = \tilde{\lambda}' \tilde{F}_t$. Thus, the factors and loadings are only identified up to a rotation and R^2 restrictions are needed to identify a certain rotation. I impose them by assuming that a submatrix of the matrix of factors is the identity matrix, which often gives the individual effects an intuitive interpretation. For example, when the outcomes are test scores, λ_1 and λ_2 can then be interpreted as the abilities, which affect test 4 and 5, respectively. Assumption L2 is a strict exogeneity assumption. Assumption L3 implies that U_t and λ are uncorrelated and that U_t is uncorrelated across t, conditional on X. Assumptions L4 and L5 ensure that the covariance matrix of any two pairs of outcomes has full rank and imply that each outcome is affected by a different linear combination of λ. Assumption L6 describes the variation in X_t over t needed to identify β_t.

Assumption L1 implies that

$$\lambda_1 = Y_4 - X_4 \beta_4 - U_4 \quad \text{and} \quad \lambda_2 = Y_5 - X_5 \beta_5 - U_5.$$

Plugging these expressions for λ into Equation (2) when $t = 3$ and rearranging yields

$$Y_3 = Y_4 F_{31} + Y_5 F_{32} + X_3 \beta_3 - X_4 \beta_4 F_{31} - X_5 \beta_5 F_{32} + \varepsilon, \quad \varepsilon = U_3 - U_4 F_{31} - U_5 F_{32}.$$

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

Clearly, Y_4 and Y_5 are correlated with ε. However, we can use (Y_1, Y_2) as instruments for (Y_4, Y_5) because (Y_1, Y_2) is uncorrelated with ε conditional on X by L2 and L3 and

$$\operatorname{cov}((Y_1, Y_2), (Y_4, Y_5) \mid X) = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix} \operatorname{cov}(\lambda \mid X),$$

which has full rank by Assumptions L4 and L5. Hence F_{31} and F_{32} are identified. Next, F_1 is identified by using Y_1, Y_4, and Y_5 to difference out λ and (Y_2, Y_3) as instruments for (Y_4, Y_5). Analogously, we can identify F_2. By Assumption L6 we can now identify β_{t_1} for all t_1 by using Y_{t_1}, Y_{t_2}, Y_{t_3} to difference out λ and the remaining outcomes as instruments. Hence, to identify all parameters we have to interchange the outcomes that serve as instruments.[9] To identify the distribution of $(U, \lambda) \mid X$, stronger assumptions are needed. In particular, we could assume that U_t is independent over t and independent of λ and then use arguments related to the extension of Kotlarski's Lemma in Evdokimov and White (2012).

The arguments can easily be extended to the case where $R > 2$ and $T > 5$. However, the previous arguments highlight that it is necessary to have $T \geq 2R+1$.[10] We need $R+1$ outcomes to difference out λ and then another R outcomes, which can be used as instruments.

2.2. *Assumptions and definitions*

I now return to the general model. One assumption I impose is that the structural functions g_t are strictly increasing in the second argument, which is common in non-additive models (see for example Matzkin (2003) or Evdokimov (2010)). In the application $\lambda' F_t + U_t$ could be interpreted as the skills needed for test t and the assumption then says that more skills increase the test scores. Define the inverse function $h_t(Y_t, X_t) \equiv g_t^{-1}(Y_t, X_t)$ so that

$$h_t(Y_t, X_t) = \lambda' F_t + U_t, \qquad t = 1, \ldots, T. \tag{3}$$

Although $T \geq 2R+1$ is needed, to simplify the notation I assume that $T = 2R+1$. The extension to a larger T is straightforward as discussed below. Moreover, this section focuses on the continuous case. Therefore, I make the following two assumptions.

Assumption N1. *R is known and $T = 2R+1$.*

Assumption N2. *$f_{Y_1, \ldots, Y_T, \lambda \mid X}(y_1, \ldots, y_T, v; x)$ is bounded on $\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_T \times \Lambda \times \mathcal{X}$ and continuous in $(y_1, \ldots, y_T, v) \in \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_T \times \Lambda$ for all $x \in \mathcal{X}$. All marginal and conditional densities are bounded.*

Let $h_t'(y_t, x_t)$ denote the derivative of h_t with respect to y_t. The next assumption imposes monotonicity and a normalization on h_t.

9. These arguments differ from Ahn *et al.* (2013), who study a linear factor model for fixed T, because I allow β_t to be time varying and I use outcomes as instruments once the individual effects are differences out.

10. It can be shown that without additional assumptions to the ones presented here, the slope coefficients β_t are not point identified if $T < 2R+1$.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

Assumption N3.

(i) h_t *is strictly increasing and differentiable in its first argument.*
(ii) There exist $\bar{x} \in \mathcal{X}$ *and* $\bar{y}$ *on the support of* $Y \mid X = \bar{x}$ *such that* $h_t(\bar{y}_t, \bar{x}_t) = 0$ *for all* $t = R+2, \ldots, 2R+1$ *and* $h'_t(\bar{y}_t, \bar{x}_t) = 1$ *for all* $t = 1, \ldots, T$.

Define the subset of the support where the location normalizations are imposed by $\tilde{\mathcal{X}} \equiv \{(x_1, \ldots, x_T) \in \mathcal{X} : x_t = \bar{x}_t \text{ for all } t = R+2, \ldots, 2R+1\}$. Next, let $F = (F_1 F_2 \cdots F_T)$ be the $R \times T$ matrix of factors and let $I_{R \times R}$ be the $R \times R$ identity matrix. The remaining assumptions are as follows.

Assumption N4. $M[U_t \mid X, \lambda] = 0$ *for all* $t = 1, \ldots, T$.

Assumption N5. $U_1, \ldots, U_T, \lambda$ *are jointly independent conditional on* X.

Assumption N6. $(F_{R+1} \cdots F_{2R+1}) = I_{R \times R}$ *and any* $R \times R$ *submatrix of* F *has full rank.*

Assumption N7. *The* $R \times R$ *covariance matrix of* λ *has full rank conditional on* X.

Assumption N8. *The characteristic function of* U_t *is non-vanishing on* $(-\infty, \infty)$ *for all* t *and* λ *has support on* $\mathbb{R}^R$ *conditional on* X.

To better understand the normalizations, notice that a special case without covariates is $\alpha_t + \beta_t Y_t = \lambda' F_t + U_t$. Since the right-hand side is not observed, one can divide both sides by a constant for each t and still satisfy all assumptions. Thus, β_t is not identified for any t and N3(ii) normalizes them to 1. Similarly, α_t is not identified for R periods, because the mean of λ is unknown, and N3(ii) normalizes them to $-\bar{y}_t$ for $t = R+2, \ldots, 2R+1$. As stated in Theorem 2, economically interesting quantities, such as average and quantile structural functions, are invariant to these normalizations (as well the ones in N4 and N6).

Assumptions N4–N7 can be seen as the non-parametric analogs of L2–L5. Assumption N4 implies that the regressors are strictly exogenous with respect to U_t, which rules out for example that X_t contains lagged dependent variables. A median normalization is more convenient in non-linear models than the zero mean assumption used in the linear model. Assumption N5 strengthens L3. Although the unobservables $\lambda' F_t + U_t$ are correlated over t, any dependence is due to λ. Autoregressive U_t are thus ruled out. A similar assumption is needed in the linear model to identify the distribution of $(U, \lambda) \mid X$. Note that the assumptions do not require that U_t and X_t are independent and permit heteroskedasticity. Independence can be relaxed if $T > 2R+1$ because the proof only requires that $2R+1$ outcomes are independent conditional on (X, λ). Hence, one could allow for $MA(1)$ disturbances if $T \geq 4R+1$ and similarly for a more complicated dependence structure for larger T. Assumption N6 generalizes L1 and L4. Assumption N7 is just as L5 and rules out that some element of λ is a linear combination of the other elements. Furthermore, all constant elements of λ, and thus time trends, are absorbed by the function h_t.

Assumption N8 is an additional assumption needed due to the non-parametric nature of the model. A non-vanishing characteristic function holds for many standard distributions such as the normal family, the t-distribution, or the gamma distribution, but not for all distributions, for instance the uniform distribution. The purpose of the assumption is to guarantee that a non-parametric analog of the rank condition holds, known as completeness, which implies a strong dependence between two vectors of outcomes, similar as in the linear model (see Lemma 1 in the Appendix).

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

2.3. *Identification outline and main results*

I now outline the main identification arguments and state and discuss the formal results. The first step is to notice that independence of $U_1,\ldots,U_T,\lambda\,|\,X$ implies that

$$f_{Y|X}(y;x)=\int\prod_{t=1}^{T}f_{Y_t|X,\lambda}(y_t;x,v)f_{\lambda|X}(v;x)dv.$$

Similarly, with $Z_1\equiv(Y_1,\ldots,Y_R)$, $Z_2\equiv Y_{R+1}$, and $Z_3\equiv(Y_{R+2},\ldots,Y_{2R+1})$ we get

$$f_{Y|X}(y;x)=\int f_{Z_1|X,\lambda}(z_1;x,v)f_{Z_2|X,\lambda}(z_2;x,v)f_{Z_3,\lambda|X}(z_3,v;x)dv. \tag{4}$$

The expression for $f_{Y|X}(y;x)$ has a similar structure as in the measurement error model of Hu and Schennach (2008). Here we can interpret λ as unobserved regressors. By Assumption N6, we can solve for λ in terms of any R outcomes, the corresponding X_t, and U_t. Thus, a set of R outcomes, here Z_3, can be interpreted as observed, but mismeasured regressors. The instruments needed for identification are then another set of R outcomes, Z_1, as before.

The results of Hu and Schennach (2008) do not immediately apply to (4) for two main reasons. First, since Z_2 is of lower dimension than λ and since I assume a factor structure for the unobservables, one of their identification conditions is violated.[11] I solve this problem by rotating the outcomes contained in Z_1 and Z_2, which is analogous to rotating the outcomes that serve as instruments in the linear model.[12] This additional step and arguments as in Hu and Schennach (2008) then imply identification of $f_{Y,\lambda|X}$ up to a one-to-one transformation of λ. Second, to pin down this transformation, Hu and Schennach (2008) and Cunha *et al*. (2010) impose a normalization of the form $\Psi(f_{Z_3|\lambda}(\cdot\,|\,\lambda))=\lambda$, where Ψ is a known functional, such as $E(Z_3\,|\,\lambda)=\lambda$ in a classical measurement error model. However, in the factor model discussed here, I show that such a normalization imposes very strong restrictions on the structural functions and that all parameters are identified without an additional normalization of λ.[13] To do so, I use arguments from linear factor models and single index models to point identify all parameters of the model. Important assumptions used in this step are the factor structure, independence, monotonicity, the normalizations of g_t, and the moments conditions. These arguments then not only uniquely determine $f_{Y,\lambda|X}$, but also g_t and F_t. To obtain these results I require stronger independence assumptions compared to Cunha *et al*. (2010), but some of these assumptions also serve as sufficient conditions for their completeness assumptions and are used to identify average and quantile structural functions. These arguments lead to the following theorem. The proof is in the appendix.

Theorem 1. *Suppose Assumptions N1 – N8 hold. Then F_t, the functions g_t, and the distribution of $(U,\lambda)\,|\,X=x$ are identified for all $x\in\tilde{\mathcal{X}}$. If in addition $f_X(x)>0$ for all $x\in\mathcal{X}_1\times\ldots\times\mathcal{X}_T$, then g_t and the distribution of $(U,\lambda)\,|\,X=x$ are identified for all $x\in\mathcal{X}$.*

11. Specifically, Assumption 4 in Hu and Schennach (2008) or, translated to the panel setting, Assumption (iv) or Theorem 2 in Cunha *et al*. (2010).

12. For this particular step, I require $U_1\perp\!\!\!\perp U_2\perp\!\!\!\perp\ldots\perp\!\!\!\perp U_{R+1}\perp\!\!\!\perp(U_{R+2},\ldots,U_{2R+1})\,|\,\lambda$ as opposed to $(U_1,U_2,\ldots\ldots U_R)\perp\!\!\!\perp U_{R+1}\perp\!\!\!\perp(U_{R+2},\ldots,U_{2R+1})\,|\,\lambda$ without rotations in Hu and Schennach (2008).

13. Only an unknown transformation of λ is pinned down by the eigenfunctions. For example, Assumptions N3(i), N4, and N6 imply that $M[Y_T\,|\,X=x,\lambda=v]=g_T(x_T,v_R)$. In the completely non-parametric setting of Hu and Schennach (2008) and Cunha *et al*. (2010), this assumption is much less restrictive and is truly a normalization if a monotonicity condition holds.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

Remark 1. *The proof proceeds in two steps. First, I condition on $x \in \tilde{\mathcal{X}}$ and I show that F_t, the functions g_t, and the conditional distribution of the unobservables are identified. Consequently, $f_{Y,\lambda|X}(y,v;x)$ is identified for all $y \in \mathcal{Y}$, $v \in \Lambda$, and $x \in \tilde{\mathcal{X}}$. The reason for conditioning on $x \in \tilde{\mathcal{X}}$ is that I make use of the normalizations in Assumption N3(ii). Notice that $\tilde{\mathcal{X}}$ is a subset of the support $\mathcal{X}$ with x_t fixed for all $t = R+2, \ldots, 2R+1$. To identify the functions g_t for different values of X_t, the covariates need to have enough variation across t, similar as in the linear model. A simple sufficient condition is that $f_X(x) > 0$ for all $x \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_T$. Section S.1.2 in the Supplementary Appendix discusses a weaker sufficient condition for the variation needed, which requires more notation, but is important for the application.*

Remark 2. *While the assumptions are natural extensions of those in the linear model, the identification arguments are different. When $T = 5$ and $R = 2$ we get just as in Section 2.1*

$$h_3(Y_3, X_3) = h_4(Y_4, X_4)F_{31} + h_5(Y_5, X_5)F_{32} + \varepsilon, \quad \varepsilon = U_3 - U_4F_{31} - U_5F_{32},$$

which might suggest that identification could be based on moment conditions. While such an approach might lead to identification of h_t under similar assumptions, my approach also yields identification of the distribution of $(\lambda, U)\,|\,X$ and thus, average and quantile structural functions, which require knowledge of the distribution of the unobservables and are invariant to the normalizing assumptions (see Section 2.4).

Remark 3. *The Supplementary Appendix contains extensions of the identification results to identification of the number of factors (Section S.1.1 in Supplementary Appendix), lagged dependent variables as regressors (Section S.1.4 in Supplementary Appendix), and discrete outcomes (Section 1.1.5 in Supplementary Appendix). Incorporating predetermined regressors other than lagged dependent variables requires modeling their dependence, similar as in Shiu and Hu (2013). Lagged dependent variables have the advantage of being modeled in the system.*

2.4. *Objects invariant to normalizations*

This section describes economically interesting objects, namely average and quantile structural functions, which are invariant to the normalization assumptions N3(ii), N4, and N6. Define $C_t \equiv \lambda' F_t$ and let $Q_\alpha[C_t]$ and $Q_\alpha[U_t]$ be the α-quantile of C_t and U_t, respectively. Let $\tilde{x}_t \in \mathcal{X}_t$ and define the quantile structural functions

$$s_{t,\alpha}(\tilde{x}_t) = g_t(\tilde{x}_t, Q_\alpha[C_t + U_t]) \quad \text{and} \quad s_{t,\alpha_1,\alpha_2}(\tilde{x}_t) = g_t\left(\tilde{x}_t, Q_{\alpha_1}[C_t] + Q_{\alpha_2}[U_t]\right)$$

as well as the average structural function $\bar{s}_t(\tilde{x}_t) = \int g_t(\tilde{x}_t, e)\, dF_{C_t+U_t}(e)$. The functions $s_{t,\alpha}(\tilde{x}_t)$ and $\bar{s}_t(\tilde{x}_t)$ are analogous to the average and quantile structural functions in Blundell and Powell (2003) and Imbens and Newey (2009). Here the unobservables consist of two parts, C_t and U_t, and C_t often has a specific interpretation in applications, such as the abilities needed for a certain test. The function $s_{t,\alpha_1,\alpha_2}(\tilde{x}_t)$ allows the two unobservables to be evaluated at different quantiles. Therefore, one could set U_t to its median value of 0 and investigate how the outcomes vary with C_t. Moreover, let $x \in \mathcal{X}$ and define the conditional versions of these functions as

$$\begin{aligned} s_{t,\alpha}(\tilde{x}_t, x) &= g_t(\tilde{x}_t, Q_\alpha[C_t + U_t \,|\, X = x]) \\ s_{t,\alpha_1,\alpha_2}(\tilde{x}_t, x) &= g_t\left(\tilde{x}_t, Q_{\alpha_1}[C_t \,|\, X = x] + Q_{\alpha_2}[U_t \,|\, X = x]\right), \text{ and} \\ \bar{s}_t(\tilde{x}_t, x) &= \int g_t(\tilde{x}_t, e)\, dF_{C_t+U_t|X=x}(e). \end{aligned}$$

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

Average and quantile structural functions can be used to answer important policy questions. For example suppose X_t is class size and the outcomes are test scores. Then $\bar{s}_t(25)$–$\bar{s}_t(20)$ is the expected effect of a change in class size from 20 to 25 on the test score for a randomly selected student. The conditional version $\bar{s}_t(25,x)$–$\bar{s}_t(20,x)$ is the expected effect for a randomly selected student from a class of size x. The quantile effects have similar interpretations, but are evaluated at quantiles of unobservables, rather than averaging over them. The following result shows identification of these functions without the normalizations.

Theorem 2. *Suppose Assumptions N1, N2, N3(i), N5, N7, and N8 hold. Further suppose that for all t $M[U_t|X,\lambda]=c_t$ for some $c_t\in\mathbb{R}$, that each $R\times R$ submatrix of F has full rank, and that $f_X(x)>0$ for all $x\in\mathcal{X}_1\times\ldots\times\mathcal{X}_T$. Then $s_{t,\alpha}(\tilde{x}_t,x)$, $s_{t,\alpha_1,\alpha_2}(\tilde{x}_t,x)$, $\bar{s}_t(\tilde{x}_t,x)$, $s_{t,\alpha}(\tilde{x}_t)$, $s_{t,\alpha_1,\alpha_2}(\tilde{x}_t)$, and $\bar{s}_t(\tilde{x}_t)$ are identified for all $\tilde{x}_t\in\mathcal{X}_t$ and $x\in\mathcal{X}$.*

Remark 4. *As in Theorem 1, we can replace $f_X(x)>0$ for all $x\in\mathcal{X}_1\times\ldots\times\mathcal{X}_T$ with a weaker sufficient condition. Specifically, we can instead assume that Assumption N9 in Section S.1.2 in the Supplementary Appendix holds for all $x\in\mathcal{X}$.*

3. ESTIMATION

This section discusses estimation when R is known. Section S.2.3 in the Supplementary Appendix shows how to test the null hypothesis that the model has R factors against the alternative that it has more than R factors, and how to consistently estimate the number of factors.

First notice that, by Assumptions N3 and N5, the density of $Y|X$ can be written as

$$f_{Y_1,\ldots,Y_T|X}(y;x)=\int\prod_{t=1}^{T}f_{U_t|X}(h_t(y_t,x_t)-v'F_t;x)h_t'(y_t,x_t)f_{\lambda|X}(v;x)dv.$$

I use this expression to suggest estimation based on the maximum likelihood method. Although I show that a completely non-parametric estimator is consistent, such an estimator might not be attractive in applications due to the potentially high dimensionality of the estimation problem. For example, the function $f_{\lambda|X}(v;x)$ has $R+Td_x$ arguments, which implies a slow rate of convergence, and consequently imprecise estimators in finite samples.[14] Hence, I also suggest a more practical semiparametric estimator, where I reduce the dimensionality by assuming a location and scale model for the conditional distributions.

3.1. *Fully non-parametric estimator*

I follow well known results, such as Chen (2007), and prove consistency of a non-parametric maximum likelihood estimator. I briefly outline the main assumptions below and provide the details in the Supplementary Appendix. Next to the identification conditions, the main assumptions for estimation include smoothness restrictions on the unknown functions. Specifically, I assume that the unknown functions lie in a weighted Hölder space, which allows the functions to be unbounded and have unbounded derivatives. I denote the parameter space by Θ and the consistency norm by $\|\cdot\|_s$, which is a weighted sup norm.[15] Let $W_i=(Y_i,X_i)$ and

14. In addition, the model nests deconvolution problems which can have a logarithmic rate of convergence. For related setups see Fan (1991), Delaigle *et al.* (2008), and Evdokimov (2010).

15. This combination of the consistency norm and the parameter space ensures that Θ is compact under $\|\cdot\|_s$. As discussed in Section S.2.1 in Supplementary Appendix, a weighted sup norm implies consistency in the regular unweighted sup norm over any compact subset of the support.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

denote the true value of the parameters by $\theta_0 = \left(h_1, \ldots, h_T, f_{U_1|X}, \ldots, f_{U_T|X}, f_{\lambda|X}, F\right) \in \Theta$. Then the log-likelihood evaluated at θ_0 and the ith observation is

$$l(\theta_0, W_i) \equiv \ln \int \prod_{t=1}^{T} f_{U_t|X}(h_t(Y_{it}, X_{it}) - v' F_t; X_{it}) h_t'(Y_{it}, X_{it}) f_{\lambda|X}(v; X_{it}) dv.$$

Now let Θ_n be a finite dimensional sieve space of Θ, which depends on the sample size n and has the property that θ_0 can be approximated arbitrary well by some element in Θ_n when n is large enough (see Assumption E4 in the Supplementary Appendix for the formal statement). For example, h_t could be approximated by a polynomial function, where the order of the polynomial grows with the sample size. The estimator of θ_0 is $\hat{\theta} \in \Theta_n$ which satisfies

$$\frac{1}{n} \sum_{i=1}^{n} l(\hat{\theta}, W_i) \geq \sup_{\theta \in \Theta_n} \frac{1}{n} \sum_{i=1}^{n} l(\theta, W_i) - o_p(1/n).$$

Once the sieve space is specified, estimation is equivalent to a parametric maximum likelihood estimator.[16] For the estimator to be consistent it is crucial that the parameter space reflects all identification assumptions to ensure that θ_0 is the unique maximizer of $E[l(\theta, W_i)]$ in Θ. Notice that the likelihood already incorporates independence of $U_1, \ldots, U_T, \lambda$. Moreover, the normalizations in Assumptions N3(ii), N4, and N6 as well as monotonicity of h_t are straightforward to impose (see Section 5 for details). The remaining two assumptions, N7 and N8, do not have to be imposed in the optimization problem. The reason is that even without imposing the assumptions, a maximizer of $E[l(\theta, W_i)]$ corresponds to the true density of $Y|X$. By Lemma 1 this density implies certain completeness conditions, which can only hold if the covariance matrix of $\lambda|X$ has full rank. Moreover, given Assumption N1–N7, completeness is sufficient for identification and therefore θ_0 is the unique maximizer of $E[l(\theta, W_i)]$. Other implementation issues, including specific sieve spaces, are discussed in Sections 4 and 5 (the application and Monte Carlo simulations, respectively) in more detail.

The following result is shown in the appendix which, given the assumptions, follows from Theorem 3.1 in combination with Condition 3.5M in Chen (2007).

Theorem 3. *Let Assumptions N1–N8 and Assumptions E7–E9 in the Supplementary Appendix hold. Let Assumption N9 in the Supplementary Appendix hold for all $x \in \mathcal{X}$. Then $\|\hat{\theta} - \theta_0\|_s \xrightarrow{p} 0$.*

Remark 5. *It is well known that if the individual fixed effects are estimated as parameters, then the maximum likelihood estimator is generally not consistent in non-linear panel data models when T is fixed (i.e. the incidental parameters problem). I circumvents this problem by not treating the fixed effects as parameters, but instead estimating the distribution of λ. The assumptions then imply that the number of parameters grows slowly with the sample size, as opposed to being of the same order as the sample size. However, I assume that λ has a smooth density, which is not required when the fixed effects are treated as parameters (but in this case the estimator would not be consistent). I therefore rule out for example that λ is discretely distributed, but I neither impose parametric assumptions on its distribution, nor on the dependence between λ and X.*

16. The definition ensures that the estimator is always well defined. If the solution to the sample optimization problem is unique, then one can simply use $\hat{\theta} = \arg\max_{\theta \in \Theta_n} \frac{1}{n} \sum_{i=1}^{n} l(\theta, W_i)$.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

Remark 6. *Consistency of $\hat{\theta}$ in the $\|\cdot\|_s$ norm implies consistency of plug-in estimators of average and quantile structural functions. For example, let $\hat{s}_{t,\alpha}(\tilde{x}_t)=\hat{g}_t(\tilde{x}_t,\hat{Q}_\alpha[C_t+U_t])$, where $\hat{g}_t$ is the estimated structural function and $\hat{Q}_\alpha[C_t+U_t]$ is the estimated α quantile of C_t+U_t obtained from the estimated density. Then the assumptions and results of Theorem 3 imply that $\hat{s}_{t,\alpha}(\tilde{x}_t)\xrightarrow{p} s_{t,\alpha}(\tilde{x}_t)$.*

3.2. *Semiparametric estimator*

I now outline a semiparametric estimator, which I use in the application. First, I reduce the dimensionality of the estimation problem by making additional assumptions on the conditional distribution of $\lambda\,|\,X$. In particular, I assume that $\lambda=\mu(X,\beta_1)+\Sigma(X,\beta_2)\varepsilon$, where ε is independent of X. The main advantage of this approach is that the likelihood now depends on the density of ε as well as β_1 and β_2 instead of the high dimensional function $f_{\lambda|X}$. Furthermore, I assume that U_t is independent of X, but the density f_{U_t} is unknown. Alternatively, one could model the dependence between U_t and X to allow for heteroskedasticity. The structural functions can be parametric, semiparametric, or non-parametric depending on the application. To accommodate all cases, I assume that $h_t(Y_t,X_t)=m(Y_t,X_t,\alpha_t,\beta_{3t})$, where β_{3t} is a finite dimensional parameter, α_t is an unknown function, and m is a known function. As an example, in Sections 4 and 5, I model $h_t(Y_t,X_t)=\alpha_t(Y_t)-X_t'\beta_{3t}$.

Define the finite dimensional parameter vector $\beta_0=(\beta_1,\beta_2,\beta_{31},\ldots,\beta_{3T},F)'$, let $\alpha_0=(\alpha_1,\ldots,\alpha_T,f_\varepsilon,f_{U_1},\ldots,f_{U_T})$ denote all unknown functions, and define $\theta_0\equiv(\alpha_0,\beta_0)$. Now in addition to the various finite dimensional parameters, several low dimensional functions, namely T one-dimensional densities, one R-dimensional density, and the functions α_t have to be estimated. The estimator $\hat{\theta}=(\hat{\alpha},\hat{\beta})$ is again computed using sieves and maximizing the log-likelihood function. This is computationally almost identical to the estimator described in the previous section, except that now there are less sieve terms and more finite dimensional parameters to maximize over. Next to improved rates of convergence due to the reduced dimensionality, another major advantage of the semiparametric estimator is that β_0 can be estimated at the $\sqrt{n}$ rate and the estimator is asymptotically normally distributed. Thus, one can easily conduct inference. These results are shown in the following theorem.

Theorem 4. *Let Assumptions E2 and E8–E18 in the Supplementary Appendix hold. Then $\sqrt{n}\big(\hat{\beta}-\beta_0\big)\xrightarrow{d} N\big(0,(V^*)^{-1}\big)$, where V^* is defined in Equation (4) in the Supplementary Appendix.*

The proof is very similar to the ones in Ai and Chen (2003) and Carroll *et al*. (2010) among others. Ackerberg *et al*. (2012) provide a consistent estimator of the covariance matrix and discuss its implementation in a more general setting.

3.3. *Parametric estimator*

Finally, given the previous results, it is straightforward to estimate the model completely parametrically. In this case the densities f_{U_t} and $f_{\lambda|X}$ and the functions h_t are assumed to be known up to finite dimensional parameters. For example, one could assume that λ and U_t are normally distributed, where the mean and the covariance of λ is a parametric function of X and the variance of U_t is a constant. Consistency and asymptotic normality then follows from standard arguments, such as those in Newey and McFadden (1994).

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

4. APPLICATION

This section investigates the relationship between teaching practice and student achievement using test scores from the Trends in International Mathematics and Science Study (TIMSS).

4.1. *Data and background*

The TIMSS is an international assessment of mathematics and science knowledge of fourth and eighth-grade students. I make use of the 2007 sample of eighth-grade students in the U.S. This sample consists of 7,377 students. Each student attends a math and an integrated science class with different teachers in each class for most students. I exclude students which cannot be linked to their teachers, students in classes with less than five students, and observations with missing values in covariates (defined below).

The TIMSS contains test scores for different cognitive domains of the tests, which are mathematics applying, knowing, and reasoning, as well as science applying, knowing, and reasoning.[17] I use these six test scores as the dependent variables Y_{it}, where i denotes a student and t denotes a test. Hence, $T=6$ which allows me to estimate a factor model with two factors. The main regressors are measures of modern and traditional teaching practice. Intuitively, modern teaching practice is associated with group work and reasoning, while traditional teaching practice is based on lectures and memorizing. To construct these, I follow Bietenbeck (2014) and use students' answers about frequencies of certain class activities. I number the response as 0 for *never*, 0.25 for *some lessons*, 0.5 to *about half of the lessons*, and 1 for *every or almost every lesson*, so that the numbers correspond approximately to the fraction of time the activities are performed in class. The teaching measures of student i are the class means of these responses, excluding the student's own response.[18]

Various educational organizations have generally advocated for a greater use of modern teaching practices and a shift away from traditional teaching methods (see Zemelman *et al.* (2012) for a "consensus on best practice" and a list of sources, including among many others, the National Research Council and the National Board of Professional Teaching Standards). However, despite these policy recommendations, the empirical evidence on the relationship between teaching practice and test scores is not conclusive and varies depending on the data set, test scores, and methods used. For example, Schwerdt and Wuppermann (2011) make use of the 2003 TIMSS data and find positive effects of traditional teaching practice. Bietenbeck (2014) documents a positive effect of traditional and modern teaching practice on applying/knowing and reasoning test scores, respectively. Using Spanish data, Hidalgo-Cabrillana and Lopez-Mayany (2015) find a positive effect of modern teaching practice on math and reading test scores and, with teaching measures constructed from students' responses, a negative effect of traditional teaching practice. Lavy (2016) finds evidence of positive effects of both modern and traditional teaching practices on test scores using data from Israel. All of these studies at most allow for an additive student individual effect. Since math includes sections on number, geometry, algebra, data, and chance and science includes biology, chemistry, earth science, and physics, it is not clear a priori that the two subjects require the same skills.[19] I show below that the conclusions in the models I estimate change considerably once more general heterogeneity is allowed for. Moreover, while

17. "Knowing" measures knowledge of facts, concepts, and procedures. "Applying" focuses on the ability of students to solve routine problems. "Reasoning" covers unfamiliar situations, complex contexts, and multi-step problems.

18. The questions used to construct the teaching practice measures are listed in Table S.2 in the Supplementary Appendix. Bietenbeck (2014) contains much more details on their construction and the background literature.

19. For example, a physics "knowing" question asks what happens to an iron nail with an insulated wire coiled around it, which is connected to a battery, when current flows through the wire (answer: the nail will become a magnet). An algebra "knowing" question asks what $\frac{x}{3}>8$ is equivalent to (answer: $x>24$).

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

Zemelman *et al*. (2012) generally advocate for modern teaching practices in all subjects, best teaching practices vary across subjects. For instance, they write that "we now know that problem solving can be a means to build mathematical knowledge" (p. 170). It is thus not obvious that the same teaching practice dominates in both subjects and I therefore also allow for different effects of teaching practices across test scores.[20]

The outcome equation of the general model is $Y_t = g_t(X_t, \lambda' F_t + U_t)$ and thus, Y_t is an unknown function of X_t. Hence, if X_t is discrete, a completely non-parametric estimator allows for a different function for each point of support of the covariates, and a researcher can study the differences of the estimated functions for different values of X_t. A major downside of this generality is that there might be very few observations once all discrete covariates are controlled for. To keep the non-parametric idea of the estimator, in this application I restrict myself to students between the age of 13.5 and 15.5 and English as their first language, which leaves 1,739 male and 1,787 female students in 169 schools with 235 math and 265 science teachers.[21] I then estimate the model separately for male and female students to illustrate how discrete covariates can be incorporated non-parametrically, and how gender heterogeneity can be studied with the non-parametric estimator. Similarly, the general model allows for a completely non-parametric function of all additional covariates, including teaching practices, but estimating functions of many dimensions implies a slow rate of convergence and poor finite sample properties. I therefore estimate a flexible semiparametric model, similar to the one in the Monte Carlo simulations, which allows among others for an unknown transformation of the test scores.

4.2. *Model and implementation*

The results reported in this article are based on the outcome equation

$$\alpha_t(Y_t) = \gamma_t + X_t^{trad}\beta_t^{trad} + X_t^{mod}\beta_t^{mod} + Z_t'\delta + \lambda' F_t + U_t, \tag{5}$$

where $t = 1, 2, 3$ are the math scores (applying, knowing, reasoning) and $t = 4, 5, 6$ are the science scores (applying, knowing, reasoning). The scalars X_t^{mod} and X_t^{trad} are the modern and traditional teaching practice measures. The vector Z_t includes the other covariates, namely the class size, hours spent in class, teacher experience, whether a teacher is certified in the field, and the gender of the teacher. I set $\lambda = \mu(X^{trad}, X^{mod}, \theta) + \varepsilon$, where $\varepsilon \perp\!\!\!\perp X^{trad}, X^{mod}, Z$ and μ is a linear function of X^{mod} and X^{trad}, and $U \perp\!\!\!\perp (\lambda, X^{trad}, X^{mod}, Z)$.[22]

I estimate marginal effects, evaluated at the median value of the observables and different quantiles of $\lambda' F_t$.[23] There are twelve marginal effects I consider, namely the effect of traditional teaching on Y_t and the effect of modern teaching on Y_t for $t = 1, \ldots, 6$, which correspond to the derivative of the quantile structural function, $s_{t,q,\frac{1}{2}}(\tilde{x}_t)$, discussed in Section 2.4. With the

20. Other settings where estimated effects differ considerably between math and science include the effects of degrees/coursework and the gender of the teacher on student achievement, respectively (see Wayne and Youngs (2003) and Dee (2007)).

21. I obtain qualitatively similar results for a smaller sample, with 897 male and 973 female students, which is restricted to schools with an enrollment between 100 and 600 students, where parents' involvement is not reported to be very low, and where less than 75% of the students receive free lunch.

22. With this assumption, λ become correlated random effects instead of fixed effects. The results with a quadratic μ are almost identical.

23. Estimation results based on the average structural functions, $\bar{s}_t(x_t)$, and averaged over the covariates, are very similar.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

specification above, the marginal effect of traditional teaching is

$$\frac{\partial}{\partial x_t^{trad}} \alpha_t^{-1}\left(\gamma_t + \tilde{x}_t^{trad}\beta_t^{trad} + \tilde{x}_t^{mod}\beta_t^{mod} + \tilde{z}_t'\delta + Q_q[\lambda' F_t]\right), \tag{6}$$

where $\tilde{x}_t^{trad} = M[X_t^{trad}]$, $\tilde{x}_t^{mod} = M[X_t^{mod}]$, and $\tilde{z}_t = M[Z_t]$. In a linear model these marginal effect are simply the slope coefficients β_t^{trad} and β_t^{mod}, and therefore do not depend on the skill level. I show results for the linear fixed effects estimator (FE), three parametric models, and a semiparametric estimator. All parametric models assume that $a_t(\cdot)$ is linear and that ε and U_t are normally distributed. I consider a one factor model where $F_t = 1$ for all t, a one factor model with time varying factors, and a two factor model to illustrate what is driving the differences between the fixed effects estimates and the semiparametric estimates. In addition, I present results for a linear fixed effects model, where the slope coefficients are identical across subjects, which is the specification of Bietenbeck (2014).

For the semiparametric estimator I estimate among others six one-dimensional functions α_t, six one-dimensional functions f_{U_t}, the two-dimensional pdf of ε, and twelve slope coefficients for teaching practices. The outcome equation is only non-parametric in Y_t because a more flexible specification with higher dimensional functions would be very imprecise with the limited sample size. While this specification is relatively simple, it keeps all important features of the model, namely the two factors and heterogeneous marginal effects, and that the results do not depend on the particular metric of the test scores. The linearity in X_t^{trad} and X_t^{mod} also has the advantage that marginal effects are non-zero if and only if the slope coefficients are non-zero. Since the estimated slope coefficients are asymptotically normally distributed, we can find significance of estimated marginal effects by testing $H_0 : \beta_t^{trad} = 0$, even in the semiparametric model, which would not be possible with a completely non-parametric function. Finally, although the model is semiparametric, the structural functions are non-parametrically identified under Assumptions N1–N8 and weak support conditions on the teaching practice measures, as discussed in Section S.1.3 in Supplementary Appendix. To calculate the standard errors for the parametric and semiparametric likelihood based estimators I use the estimated outer product form of the covariance matrix as suggested by Ackerberg *et al.* (2012). For the linear fixed effects model I use standard GMM-type standard errors. I defer specific implementation issues, such as the choices of basis functions and how the constraints are imposed, to Section 5 as well as Section S.3 in the Supplementary Appendix.

4.3. *Results*

Table 1 shows the estimated marginal effects for the sample of 1,739 boys.[24] The results from the linear fixed effects models suggest a positive relationship between X_t^{trad} and knowing and applying test scores as well as a positive relationship between X_t^{mod} and reasoning scores. In the unrestricted model, the slope coefficients are similar for math and science and thus, restricting the slope coefficients to be the same across subjects yields similar results. I standardized Y_t and the teaching practice measures to have a standard deviation of 1. Hence, a one standard deviation increase of X_2^{trad} is associated with a 0.078 standard deviation increase of Y_2 in the unrestricted fixed effects model. The marginal effects for a parametric one factor model with $F_t = 1$, where α_t is linear and all unobservables are normally distributed, are very similar to the fixed effects model,

24. For each student and test, the TIMSS contains five imputed values because students generally did not answer the same set of questions. My results are based on the first imputed values for each student and test, but the results with the others are similar.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

TABLE 1
Marginal effects teaching practice for boys

		Fixed effects		Parametric			Semip.
Subject	Teaching	Restr.	Unrestr.	$R=1$ $F_t=1$	$R=1$	$R=2$	$R=2$
Math applying	Trad.	0.034*	0.041**	0.042	0.105***	0.138	0.139
Math knowing	Trad.	0.063***	0.078***	0.079**	0.142***	0.171**	0.174**
Math reasoning	Trad.	0.021	0.015	0.011	0.089***	0.117	0.120
Science applying	Trad.	0.034*	0.030	0.033***	0.068***	−0.186	−0.193
Science knowing	Trad.	0.063***	0.038*	0.035***	0.069***	−0.189	−0.198
Science reasoning	Trad.	0.021	0.029	0.031***	0.065***	−0.165	−0.173
Math applying	Modern	0.012	0.023	0.022	−0.010	−0.200**	−0.200**
Math knowing	Modern	−0.011	−0.013	−0.007	−0.039	−0.214**	−0.215**
Math reasoning	Modern	0.046**	0.049**	0.045	0.002	−0.155**	−0.159*
Science applying	Modern	0.012	0.009	0.009	0.002	0.405*	0.411*
Science knowing	Modern	−0.011	0.011	0.016*	0.009	0.421**	0.428**
Science reasoning	Modern	0.046**	0.045**	0.042***	0.035***	0.396**	0.402**

The symbols *, **, and *** denote significance at 10%, 5%, and 1% level, respectively. Significance levels are obtained by testing $H_0:\beta_t^{trad}=0$ and $H_0:\beta_t^{mod}=0$.

which is not surprising because they are based on the same outcome equation. However, in the fixed effects model, U_t is not assumed to be independent over t and the relation between λ and X is not modeled. Independence might be hard to justify here because all three math (and similarly science) test scores are obtained from the same overall test. Nonetheless, the two models yield very similar conclusions. Allowing F_t to vary produces different marginal effects, which now suggest that traditional teaching practices are associated with better test scores in both subjects. Moreover, in this model $\hat{\beta}_t^{trad} > \hat{\beta}_t^{mod}$ for all t.

Allowing for two individual effects changes the estimates considerably. Specifically, a parametric two factor model still yields a positive relationship between X_t^{trad} and math scores, but a negative relationship between X_t^{trad} and science scores. Contrarily, X_t^{mod} has a positive effect on science and a negative effect on math. The effect of X_t^{trad} on math knowing scores and the effects of X_t^{mod} on all tests are significantly different from 0. Furthermore, I reject $H_0:\beta_1^{trad}=\beta_2^{trad}=\beta_3^{trad}=0$ and $H_0:\beta_1^{mod}=\beta_2^{mod}=\beta_3^{mod}=0$ at the 1% level and $H_0:\beta_4^{mod}=\beta_5^{mod}=\beta_6^{mod}=0$ at the 2% level. For modern teaching practice I also reject that the marginal effects in the two factor model are the same as the ones in the linear fixed effects model (for each t at least at the 10% level). The estimated matrix of factors is

$$\begin{array}{c} \\ \text{Skill 1} \\ \text{Skill 2} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{Math applying} & \text{Math knowing} & \text{Math reasoning} & \text{Science applying} & \text{Science knowing} & \text{Science reasoning} \end{array} \\ \begin{pmatrix} 1.00 & 0.94 & 0.84 & 0.03 & 0.00 & 0.11 \\ 0.00 & 0.04 & 0.03 & 0.98 & 1.00 & 0.89 \end{pmatrix} \end{array}.$$

The math subjects have more weight on the first skill, while science subjects have more weight on the second skill. Two numbers are exactly 0 and two are exactly 1, which corresponds to a particular normalization. That is, λ_1 can be interpreted as the skills needed for math applying and λ_2 are the skills for science knowing. Hence, the skills needed in other subjects are linear combinations of these two skills. The estimated correlation is around 68%. Notice that identification would fail if two factors, next to F_{12} and F_{51}, were zero. Using the results in Chen *et al*. (2011), I can test whether any combination of two factors are 0 and I reject each such null at least at the 10% level. I also reject the one factor model in favour of the two factor model at

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

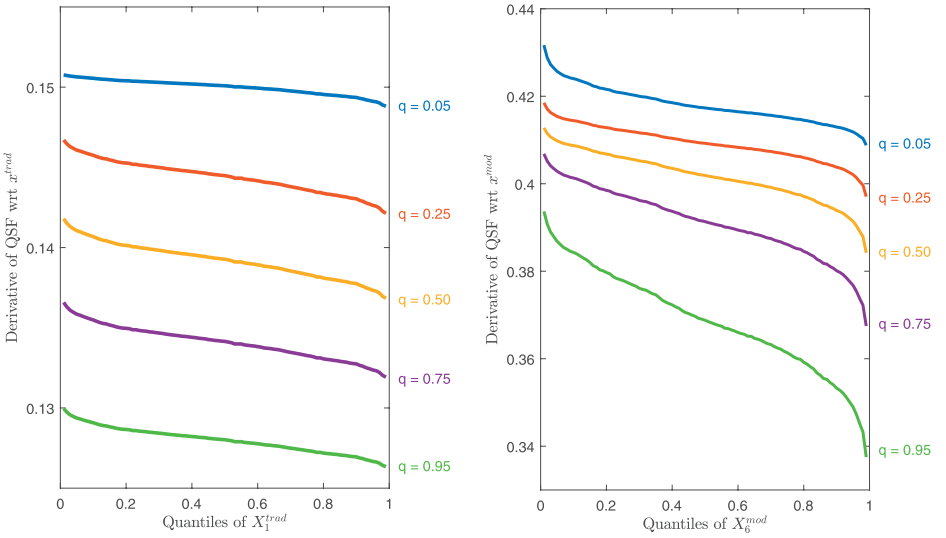

FIGURE 1
Derivatives of quantile structural functions

the 1% level. The Appendix also contains results for the sample of 1,787 girls. While the results are mostly qualitatively similar, the estimated marginal effects of tradition teaching practices on math scores are not statistically significant and negative, suggesting heterogeneity in gender.

The estimated marginal effects in the semiparametric model, evaluated at the median of the observables and unobservables, are very similar to the ones in the parametric two factor model. The additional conclusions one can draw from a non-linear model are illustrated in Figure 1, which shows derivatives of quantile structural functions, namely estimates of

$$\frac{\partial}{\partial x_1^{trad}} \alpha_1^{-1}\left(\gamma_1 + x_1^{trad}\beta_1^{trad} + \tilde{x}_1^{mod}\beta_1^{mod} + \tilde{z}_1'\delta + Q_q[\lambda' F_1]\right)$$

in the left panel (as a function of quantiles of X_1^{trad} and for different quantiles of $\lambda' F_1$) and

$$\frac{\partial}{\partial x_6^{mod}} \alpha_6^{-1}\left(\gamma_6 + \tilde{x}_6^{trad}\beta_6^{trad} + x_6^{mod}\beta_6^{mod} + \tilde{z}_6'\delta + Q_q[\lambda' F_6]\right)$$

in the right panel (as a function of quantiles of X_6^{mod} and for different quantiles of $\lambda' F_1$).[25] The results suggest that marginal effects are larger for small values of teaching practices and larger for students with low abilities, because the smaller q, the larger the function values. Hence, changes in teaching practices seem to have a larger impact on low ability students. These conclusions generally also hold for the other ten marginal effects as shown in Table 2. This table displays derivatives of the quantile structural functions for different quantiles of $\lambda' F_t$ (high skills is the 95% quantile, medium the 50% quantile, and low skills the 5% quantile) and evaluated at the median values of the covariates. Similar as in Figure 1, the marginal effects are usually largest in absolute value for students with low abilities.

25. Using quantiles of $\lambda' F_t + U_t$ yields similar results and even more heterogeneity due to the presence of the additional random variable U_t.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

TABLE 2
Marginal effects for boys and different skills

Subject	Teaching	Low skills	Medium skills	High skills
Math applying	Trad.	0.150	0.139	0.128
Math knowing	Trad.	0.174	0.174	0.165
Math reasoning	Trad.	0.118	0.120	0.109
Science applying	Trad.	−0.197	−0.193	−0.183
Science knowing	Trad.	−0.202	−0.198	−0.185
Science reasoning	Trad.	−0.181	−0.173	−0.157
Math applying	Modern	−0.215	−0.200	−0.183
Math knowing	Modern	−0.216	−0.215	−0.204
Math reasoning	Modern	−0.156	−0.159	−0.144
Science applying	Modern	0.421	0.411	0.391
Science knowing	Modern	0.436	0.428	0.400
Science reasoning	Modern	0.420	0.402	0.364

To better understand the differences between the fixed effects and the two factor model, suppose α_t is linear and suppress Z_t. Then differencing two outcomes for $t \in \{1,2,3\}$ and $s \in \{4,5,6\}$ yields

$$Y_t - Y_s = \gamma_t - \gamma_s + X_t^{trad}\beta_t^{trad} - X_s^{trad}\beta_s^{trad} + X_t^{mod}\beta_t^{mod} - X_s^{mod}\beta_s^{mod} + \lambda'(F_t - F_s) + U_t - U_s$$

and $\lambda'(F_t - F_s) = \lambda_1'(F_{t1} - F_{s1}) + \lambda_2'(F_{t2} - F_{s2})$. In this case $(F_{t1} - F_{s1}) > 0$ while $(F_{t2} - F_{s2}) < 0$, differencing might not eliminate the bias, and the direction of the bias depends on the correlation between λ and the regressors. The signs of the marginal effect changes in two cases, namely the effect of X_t^{mod} on math and X_s^{trad} on science, respectively. In the two factor model, X_t^{mod} is positively correlated with the first skill (representing applying-math) and negatively correlated with the second skill (representing knowing-science). Hence, the fixed effects model leads to a positive bias of the effect of X_t^{mod} on math, which explains the first sign change. Similarly, X_s^{trad} is negatively correlated with the first skill and positively correlated with the second skill, leading to a positive bias of the effect of X_s^{trad} on science. These correlations could either be due to teachers adapting their teaching style to the skills of the students or due to students selecting certain teachers based on their skills. Therefore, a linear fixed effects model can lead to very different conclusions compared to a model that allows for richer heterogeneity.

5. MONTE CARLO SIMULATIONS

In this section, I investigate the finite sample properties of the estimators in a setting that is calibrated to mimic the data in the empirical application. Again I let

$$\alpha_t(Y_t) = \gamma_t + X_t^{trad}\beta_t^{trad} + X_t^{mod}\beta_t^{mod} + \lambda' F_t + U_t,$$

where $X_t^{trad}, X_t^{mod} \in \mathbb{R}$, $\lambda \in \mathbb{R}^2$, and $T = 6$. Moreover, $X_t^{trad} = X_1^{trad}$ for all $t = 1,2,3$ and $X_t^{trad} = X_4^{trad}$ for all $t = 4,5,6$. The same holds for X_t^{mod}. I draw X_t^{trad} and X_t^{mod} from the empirical distribution of teaching practices I use in the application.[26] The sample size is

26. The regressors X_t^{trad} and X_t^{mod} correspond to the traditional and modern teaching practice measure, respectively. In the application $t = 1,2,3$ belongs to mathematics and $t = 4,5,6$ belongs to science test scores. Non-parametric

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

$n=1739$ as in the application. I set $\beta^{trad}=(0.14\,0.17\,0.12\,-0.19\,-0.19\,-0.17)$, and $\beta^{mod}=(-0.20\,-0.21\,-0.16\,0.41\,0.42\,0.40)$, which are the point estimates from the two factor model in the empirical application. I assume that $\lambda=\mu(X^{trad},X^{mod},\theta)+\varepsilon$, where $\varepsilon\,|\,X^{trad},X^{mod}\sim N(0,\Sigma)$ with $\Sigma_{11}=0.90$, $\Sigma_{22}=0.89$, $\Sigma_{21}=\Sigma_{12}=0.61$, and that $\mu(X^{trad},X^{mod},\theta)$ is a linear function of X_1^{trad}, X_4^{trad}, X_1^{mod}, and X_4^{mod}. Notice that the correlation between the two skills is roughly 0.68. The values of θ are also set to the point estimates and so is

$$F=\begin{pmatrix}1.00 & 0.94 & 0.84 & 0.03 & 0.00 & 0.11\\ 0.00 & 0.04 & 0.03 & 0.98 & 1.00 & 0.89\end{pmatrix}.$$

I assume that $U_t\sim N\left(0,\sigma_t^2\right)$, where $\sigma=(0.16\,0.22\,0.53\,0.21\,0.21\,0.31)$ are again the point estimates in the application. Finally, I choose $\alpha_t(Y_t)=(Y_t+c_t)^{a_t}/s_t$, where a_t, c_t, and s_t are chosen to mimic the non-parametrically estimated transformations in the application and to ensure that $\alpha_t(Y_t)$ satisfies the normalization $\alpha_t'(0)=1$. Here $a_t>1$ for all t, which implies that $\alpha_t(Y_t)$ is convex, just as the estimated functions in the empirical application.

I use five different estimators, which I also used in the empirical application, namely a linear fixed effects estimator (FE), three parametric estimators, and a semiparametric estimator. Again, all parametric estimators assume that $a_t(\cdot)$ is linear and that ε and U_t are normally distributed. The parametric estimators include a one factor model where $F_t=1$ for all t, a one factor model with time varying factors, and a two factor model. For the semiparametric estimator I non-parametrically estimate α_t, f_{U_t}, and the two-dimensional pdf of ε next to the finite dimensional parameters. To implement the semiparametric estimator, I approximate $\sqrt{f_{U_t}(u)}$ by a Hermite polynomial of degree 4, which implies that

$$f_{U_t}(u)\approx\frac{1}{\sigma_t}\left(\sum_{k=1}^{4}d_{kt}u^{k-1}\phi(u/\sigma_t)\right)^2=\frac{1}{\sigma_t}\sum_{j=1}^{4}\sum_{k=1}^{4}d_{jt}d_{kt}u^{j-1}u^{k-1}\phi(u/\sigma_t)^2,$$

where $\phi(u)$ denotes the standard normal pdf. While the theoretical arguments would allow setting $\sigma_t=1$ for all t, choosing σ_t to be an estimated standard deviation of U_t improves the finite sample properties (see Gallant and Nychka (1987) and Newey and Powell (2003) for related arguments). I set σ_t to the estimated standard deviation obtained from a parametric model. Notice that the estimated density is positive by construction. Moreover, since

$$\frac{1}{\sigma_t}\int_{-\infty}^{z}\sum_{j=1}^{4}\sum_{k=1}^{4}d_{jt}d_{kt}u^{j-1}u^{k-1}\phi(u/\sigma_t)^2du=\sum_{j=1}^{4}\sum_{k=1}^{4}d_{jt}d_{kt}\int_{-\infty}^{z/\sigma_t}u^{j-1}u^{k-1}\phi(u)^2du,$$

both the constraint that the density integrates to 1 (with $z=\infty$) and the median 0 restriction (with $z=0$) are quadratic constraints in d_{jt}. Similarly, I write $\lambda=\mu(X^{trad},X^{mod},\theta)+\Sigma^{1/2}\tilde{\varepsilon}$, I set Σ to the estimated covariance matrix from a parametric model, and I approximate the density of $\tilde{\varepsilon}$ by

$$f_{\tilde{\varepsilon}}(e_1,e_2)\approx\left(\sum_{j,k\in\mathbb{Z}^+:j+k\leq4}a_{jk}e_1^{j-1}e_2^{k-1}\phi(e_1)\phi(e_2)\right)^2.$$

identification in this setup is shown in Section S.1.3 in Supplementary Appendix. Drawing X_t^{trad} and X_t^{mod} from truncated normal distributions with the means, the covariance matrix, and the cutoffs chosen such that the distributions closely mimic the empirical distributions, yields almost identical results.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

TABLE 3
Median of estimated marginal effects and MSE

Subject	Teaching	True	FE	Parametric			Semip.
				$R=1$ $F_t=1$	$R=1$	$R=2$	$R=2$
Math applying	Trad.	0.136	0.043	0.043	0.105	0.139	0.140
			(0.009)	(0.009)	(0.001)	(0.003)	(0.003)
Math knowing	Trad.	0.170	0.080	0.078	0.140	0.173	0.173
			(0.008)	(0.008)	(0.001)	(0.003)	(0.003)
Math reasoning	Trad.	0.116	0.006	0.006	0.088	0.117	0.118
			(0.012)	(0.012)	(0.001)	(0.002)	(0.002)
Science applying	Trad.	−0.186	0.030	0.031	0.068	−0.175	−0.176
			(0.047)	(0.047)	(0.065)	(0.016)	(0.015)
Science knowing	Trad.	−0.189	0.033	0.032	0.069	−0.178	−0.179
			(0.049)	(0.049)	(0.067)	(0.017)	(0.017)
Science reasoning	Trad.	−0.163	0.029	0.031	0.065	−0.156	−0.157
			(0.037)	(0.038)	(0.052)	(0.013)	(0.013)
Math applying	Modern	−0.197	0.023	0.025	−0.009	−0.196	−0.194
			(0.048)	(0.049)	(0.035)	(0.003)	(0.003)
Math knowing	Modern	−0.213	0.000	0.000	−0.035	−0.212	−0.210
			(0.045)	(0.045)	(0.032)	(0.003)	(0.003)
Math reasoning	Modern	−0.154	0.047	0.050	0.005	−0.152	−0.150
			(0.041)	(0.042)	(0.025)	(0.002)	(0.002)
Science applying	Modern	0.403	0.015	0.014	0.004	0.386	0.385
			(0.151)	(0.151)	(0.159)	(0.022)	(0.021)
Science knowing	Modern	0.420	0.023	0.022	0.013	0.405	0.402
			(0.157)	(0.158)	(0.166)	(0.023)	(0.022)
Science reasoning	Modern	0.390	0.040	0.041	0.031	0.378	0.379
			(0.122)	(0.122)	(0.129)	(0.018)	(0.017)

The sum includes all basis functions of the form $e_1^{j-1}e_2^{k-1}\phi(e_1)\phi(e_2)$ with $j+k\leq 4$ and $j,k\geq 1$.[27] Notice that without the scale and location model, the density of $\lambda\,|\,X^{trad},X^{mod}$ would be a six-dimensional function, which would lead to imprecise estimates with a sample size of 1739. I approximate $\alpha_t(Y_t)$ with polynomials of order 4, that is $\alpha_t(Y_t)\approx Y_t+\sum_{j=2}^{4}Y_t^j b_{jt}$. The coefficient in front of Y_t is 1 to impose the scale normalization $\alpha_t'(0)=1$ and to ensure that the semiparametric model nests the linear model. The location normalizations are easy to impose by setting $\gamma_t=0$ for two periods, or by imposing $M[\lambda_j]=0$ for $j=1,2$. I use the latter restriction to facilitate comparison with a parametric model, where λ is normally distributed and $M[\lambda_j]=0$. I approximate the integral in the likelihood using Gauss-Hermite quadrature. With these choices, estimating the parameters amounts to maximizing a non-linear function subject to quadratic constraints. In Section S.3 of the Supplementary Appendix, I provide details on the convergence behavior in the simulations and the application.

I investigate finite sample properties of estimated marginal effects, evaluated at the median value of the observables and unobservables, as well as coverage rates of confidence intervals for the slope coefficients.[28] The marginal effects are analogous to those in Table 1 are described in Equation (6). The results are based on 1,000 Monte Carlo simulations. Table 3 shows the true

27. For a given number of parameters, this specification typically leads to a better approximation of the function than a tensor product (Judd, 1998).

28. While the marginal effects are invariant to the normalizations, the slope coefficients depend on the scale normalizations. Hence, imposing the true normalizations is crucial for obtaining correct coverage. In the application, I

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

TABLE 4
Coverage rates of confidence intervals with nominal level 95%

Subject	Teaching	FE	Parametric			Semip.
			$R=1$ $F_t=1$	$R=1$	$R=2$	$R=2$
Math applying	Trad.	0.001	0.201	0.995	0.958	0.966
Math knowing	Trad.	0.001	0.159	0.998	0.959	0.964
Math reasoning	Trad.	0.000	0.001	0.883	0.957	0.967
Science applying	Trad.	0.000	0.000	0.000	0.952	0.964
Science knowing	Trad.	0.000	0.000	0.000	0.957	0.965
Science reasoning	Trad.	0.000	0.000	0.000	0.955	0.965
Math applying	Modern	0.000	0.000	0.000	0.957	0.961
Math knowing	Modern	0.000	0.000	0.000	0.952	0.958
Math reasoning	Modern	0.000	0.000	0.000	0.953	0.960
Science applying	Modern	0.000	0.000	0.000	0.941	0.950
Science knowing	Modern	0.000	0.000	0.000	0.940	0.948
Science reasoning	Modern	0.000	0.000	0.000	0.939	0.953

marginal effects as well as the median of the estimated marginal effects and the median squared error (MSE) in parenthesis.[29] The fixed effects estimator and the one factor model with $F_t=1$ perform very similar and have large biases and MSEs. Time varying F_t only help reducing the biases for $t=1,2,3$. Both the parametric and the semiparametric two factor models perform very well and very similar, both in terms of the median estimated marginal effect and the MSE. The parametric model is misspecified because it assumes a linear transformation, but this seems to be a good approximation for marginal effects at the median. However, at different quantiles, the model predicts the same marginal effects, which will lead to a bias.

Table 4 shows coverage rates of confidence intervals for the estimated slope coefficients. As expected, all one factor models have poor coverage rates. Contrarily, both two factor models have coverage rates close to 95% for all slope coefficients.

6. CONCLUSION

This article studies a class of non-parametric panel data models with multidimensional, unobserved individual effects, which can impact outcomes Y_t differently for different t. These models are appealing in a variety of empirical applications where unobserved heterogeneity is not believed to be one dimensional and time homogeneous, and a researcher wants to allow for a flexible relationship between Y_t, X_t, and the unobservables. In microeconomic applications, researchers routinely use panel data to control for "abilities" using a fixed effects approach. The methods presented here allow researchers to specify much more general and realistic unobserved heterogeneity by exploiting rich enough data sets. For example, in an empirical application, I investigate the relationship between teaching practice and math and science test scores. As opposed to a standard linear fixed effects model, I allow students to have two unobserved individual effects, which can have different impacts on different tests. Hence, some students can have abilities such that they are better in math, while others can be better in science. The

am interested in testing $H_0: \beta_t^{trad}=0$, which is invariant to the normalizations and thus, coverage rates of confidence intervals for the (possibly scaled) slope coefficients are of interest.

29. I use the median and the median squared error to make the results less dependent on outliers.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

results from this model differ considerably from the ones obtained with a linear fixed effects model, which has also been used in related contexts, such as studying the relationship between student achievement and the gender of the teacher, teacher credentials, or teaching practice, respectively. Since one-dimensional heterogeneity appears to be very restrictive in this context and the conclusions from the two factor model are substantially different, specifying the most realistic model is crucial and might warrant a more in depth analysis, possibly with an even richer data set. Moreover, the models allow for heterogeneous marginal effects and thus, the effects of teaching practices on test scores can depend on students' abilities. I find that the marginal effects of a change in teaching practice on test scores are larger for students with low abilities. Next to microeconomic applications and the examples mentioned in the introduction, the models can for example also be useful in empirical asset pricing, where the return of firm i in time t, denoted by Y_{it}, can then depend on characteristics X_{it} and a small number of factors. The non-parametric approach reduces concerns about functional form misspecification (Fama and French, 2008).

I present non-parametric point identification conditions for all parameters of the models, which include the structural functions, the number of factors, the factors themselves, and the distributions of the unobservables, λ and U_t, conditional on the regressors. I also provide a non-parametric maximum likelihood estimator, which allows estimating the parameters consistently, as well as flexible semiparametric and parametric estimators.

One restriction of the models is that, other than lagged dependent variables studied in Section S.1.4 in Supplementary Appendix, the regressors are strictly exogenous. It would therefore be useful to incorporate predetermined regressors, which might require modeling their dependence. Furthermore, while Section S.2.3 in the Supplementary Appendix suggests an approach to estimate the number of factors consistently, providing an estimator with desirable finite sample properties, similar to the ones proposed by Bai and Ng (2002) in linear factor models, is another open problem. Finally, it would be interesting to extend the analysis to a large n and large T framework, where so far the existing models do not allow for interactions between covariates and unobservables.

APPENDIX

A. IDENTIFICATION PROOFS

A.1. *A useful lemma*

Lemma 1. *Let Assumptions N1, N2, N3(i), N5 – N8 hold. Let* $Z_3=(Y_{R+2},\dots,Y_{2R+1})$. *Let* $K\equiv\{k_1,k_2,\dots,k_R\}$ *be a set of any R distinct integers between* 1 *and* $R+1$. *Define* $Z_1^K\equiv\left(Y_{k_1},\dots,Y_{k_R}\right)$. *Then* Z_3 *is bounded complete for* Z_1^K *and* λ *is bounded complete for* Z_3 *given* X.

Proof. Condition on $X\in\mathcal{X}$ and suppress X. Since Z_3 and Z_1^K are independent conditional on λ,

$$f_{Z_1^K,Z_3}(z_1,z_3)=\int f_{Z_3|\lambda}(z_3;v)f_{Z_1^K|\lambda}(z_1;v)f_\lambda(v)dv.$$

It follows that for any bounded function m such that $E[|m(Z_3)|]<\infty$

$$\int f_{Z_1^K,Z_3}(z_1,z_3)m(z_3)dz_3=\int f_{Z_1^K,\lambda}(z_1,v)\left(\int f_{Z_3|\lambda}(z_3;v)m(z_3)dz_3\right)dv.$$

Conditional on $X=x$ we can write $Z_3=g(x,\lambda+V)$, where $V=(U_{R+2},\dots,U_{2R+1})$ and $g:\mathbb{R}^R\to\mathbb{R}^R$ with $g(x,v)=(g_{R+2}(x_{R+2},v_{R+2}),\dots,g_{2R+1}(x_{2R+1},v_{2R+1}))$. From Theorem 2.1 in D'Haultfoeuille (2011) it follows that Z_3 is bounded complete for λ. Furthermore, Proposition 2.4 in D'Haultfoeuille (2011) implies that λ is (bounded) complete for Z_1^K and that λ is (bounded) complete for Z_3. Hence, by the previous equality, Z_3 is bounded complete for Z_1^K. ‖

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

A.2. *Proof of Theorem 1*

First define $Z_1 \equiv (Y_1, \ldots, Y_R)$, $Z_2 \equiv Y_{R+1}$, and $Z_3 \equiv (Y_{R+2}, \ldots, Y_{2R+1})$, and let $\mathcal{Z}_1 \subseteq \mathbb{R}^R$, $\mathcal{Z}_2 \subseteq \mathbb{R}$, and $\mathcal{Z}_3 \subseteq \mathbb{R}^R$ be the supports of Z_1, Z_2, and Z_3, respectively. Next define the function spaces $\mathcal{L}^R = \{m : \mathbb{R}^R \to \mathbb{R} : \int_{\mathbb{R}^R} |m(v)| dv < \infty\}$, $\mathcal{L}^R_{bnd} = \{m \in \mathcal{L}^R : \sup_{v \in \mathbb{R}^R} |m(v)| < \infty\}$, $\mathcal{L}^R(\mathcal{Z}_1) \equiv \{m : \mathbb{R}^R \to \mathbb{R} : \int_{\mathbb{R}^R} |m(v)| f_{Z_1}(v) dv < \infty\}$ and $\mathcal{L}^R_{bnd}(\mathcal{Z}_1) \equiv \{m \in \mathcal{L}^R(\mathcal{Z}_1) : \sup_{v \in \mathbb{R}^R} |m(v)| < \infty\}$. Define $\mathcal{L}^R(\mathcal{Z}_3)$, $\mathcal{L}^R_{bnd}(\mathcal{Z}_3)$, $\mathcal{L}^R(\Lambda)$, and $\mathcal{L}^R_{bnd}(\Lambda)$ analogously. Now condition on $X = x$, where $x \in \mathcal{X}$ such that $x_t = \bar{x}_t$ for all $t = R+2, \ldots, 2R+1$, let $z_2 \in \mathbb{R}$ be a fixed constant, and define

$$L_{1,2,3} : \mathcal{L}^R_{bnd}(\mathcal{Z}_1) \to \mathcal{L}^R_{bnd} \quad (L_{1,2,3}m)(z_2, z_3) \equiv \int f_{Z_1, Z_2, Z_3|X}(z_1, z_2, z_3; x) m(z_1) dz_1$$

$$L_{1,3} : \mathcal{L}^R_{bnd}(\mathcal{Z}_1) \to \mathcal{L}^R_{bnd} \quad (L_{1,3}m)(z_3) \equiv \int f_{Z_1, Z_3|X}(z_1, z_3; x) m(z_1) dz_1$$

$$L_{3,\lambda} : \mathcal{L}^R_{bnd} \to \mathcal{L}^R_{bnd} \quad (L_{3,\lambda}m)(z_3) \equiv \int f_{Z_3|\lambda,X}(z_3; v, x) m(v) dv$$

$$L_{\lambda,1} : \mathcal{L}^R_{bnd}(\mathcal{Z}_1) \to \mathcal{L}^R_{bnd} \quad (L_{\lambda,1}m)(v) \equiv \int f_{Z_1, \lambda|X}(z_1, v; x) m(z_1) dz_1$$

$$D_{2,\lambda} : \mathcal{L}^R_{bnd}(\Lambda) \to \mathcal{L}^R_{bnd}(\Lambda) \quad (D_{2,\lambda}m)(z_2, v) \equiv f_{Z_2|\lambda,X}(z_2; v, x) m(v).$$

The operator $L_{1,2,3}$ is a mapping from $\mathcal{L}^R_{bnd}(\mathcal{Z}_1)$ to $\mathcal{L}^R_{bnd}$ for a fixed value z_2. Changing the value of z_2 gives a different mapping. With these definitions it follows from Assumption N5 that for any $m \in \mathcal{L}^R_{bnd}(\mathcal{Z}_1)$

$$\begin{aligned}
(L_{1,2,3}m)(z_2, z_3) &= \int f_{Z_1, Z_2, Z_3|X}(z_1, z_2, z_3; x) m(z_1) dz_1 \\
&= \int \left(\int f_{Z_3|\lambda,X}(z_3; v, x) f_{Z_2|\lambda,X}(z_2; v, x) f_{Z_1,\lambda|X}(z_1, v; x) dv \right) m(z_1) dz_1 \\
&= \int f_{Z_3|\lambda,X}(z_3; v, x) f_{Z_2|\lambda,X}(z_2; v, x) (L_{\lambda,1}m)(v) dv \\
&= \int f_{Z_3|\lambda,X}(z_3; v, x) (D_{2,\lambda}L_{\lambda,1}m)(z_2, v) dv \\
&= (L_{3,\lambda}D_{2,\lambda}L_{\lambda,1}m)(z_2, z_3).
\end{aligned}$$

Similarly, $(L_{1,3}m)(z_3) = (L_{3,\lambda}L_{\lambda,1}m)(z_3)$. These equalities hold for all functions $m \in \mathcal{L}^R_{bnd}(\mathcal{Z}_1)$ and thus we can write $L_{1,2,3} = L_{3,\lambda}D_{2,\lambda}L_{\lambda,1}$ and $L_{1,3} = L_{3,\lambda}L_{\lambda,1}$. By Lemma 1, $L_{3,\lambda}$ is invertible and the inverse can be applied from the left. It follows that $L^{-1}_{3,\lambda}L_{1,3} = L_{\lambda,1}$, which implies that $L_{1,2,3} = L_{3,\lambda}D_{2,\lambda}L^{-1}_{3,\lambda}L_{1,3}$. Lemma 1 of Hu and Schennach (2008) and Lemma 1 above imply that $L_{1,3}$ has a right inverse which is densely defined on $\mathcal{L}^R_{bnd}$. Therefore,

$$L_{1,2,3}L^{-1}_{1,3} = L_{3,\lambda}D_{2,\lambda}L^{-1}_{3,\lambda}.$$

The operator on the left-hand side depends on the population distribution of the observables only. Hence, it can be considered known. Hu and Schennach (2008) deal with the same type of operator equality in a measurement error setup. They show that the operator on the left-hand side is bounded and its domain can therefore be extended to $\mathcal{L}^R_{bnd}$. They also show that the right-hand side is an eigenvalue-eigenfunction decomposition of the known operator $L_{1,2,3}L^{-1}_{1,3}$. The eigenfunctions are $f_{Z_3|\lambda,X}(z_3; v, x)$ with corresponding eigenvalues $f_{Z_2|\lambda,X}(z_2; v, x)$. Each v indexes an eigenfunction and an eigenvalue. The eigenfunctions are functions of z_3, while x and z_2 are fixed. Hu and Schennach (2008) show that this decomposition is unique up to three features:

(1) Scaling: Multiplying each eigenfunction by a constant yields a different eigenvalue-eigenfunction decomposition belonging to the same operator $L_{1,2,3}L^{-1}_{1,3}$.

(2) Eigenvalue degeneracy: If two or more eigenfunctions share the same eigenvalue, any linear combination of these functions are also eigenfunctions. Then several different eigenvalue-eigenfunction decompositions belong to the same operator $L_{1,2,3}L^{-1}_{1,3}$.

(3) Ordering: Let $\tilde{\lambda} = B(\lambda, x)$ for any one-to-one transformation $B : \mathbb{R}^R \to \mathbb{R}^R$. Then $L_{3,\lambda}D_{2,\lambda}L^{-1}_{3,\lambda} = L_{3,\tilde{\lambda}}D_{2,\tilde{\lambda}}L^{-1}_{3,\tilde{\lambda}}$.

These conditions are very similar to conditions for non-uniqueness of an eigendecomposition of a square matrix. While for matrices the order of the columns of the matrix that contains the eigenvectors is not fixed, with operators any one-to-one transformation of λ leads to an eigendecomposition with the same eigenvalues and eigenfunctions (but in a different order). I show next that the assumptions fix the scaling and the ordering and that all eigenvalues are unique. It then follows that there are unique operators $L_{3,\lambda}$ and $D_{2,\lambda}$ such that $L_{1,2,3}L^{-1}_{1,3} = L_{3,\lambda}D_{2,\lambda}L^{-1}_{3,\lambda}$.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

First, the *scale* of the eigenfunctions is fixed because the eigenfunctions we are interested in are densities and therefore have to integrate to 1. Second, two different eigenfunctions share the same eigenvalue if there exists v and w with $v \neq w$ such that $f_{Z_2|\lambda,X}(z_2;v,x) = f_{Z_2|\lambda,X}(z_2;w,x)$. Following Hu and Schennach (2008), while this could happen for a fixed z_2, changing z_2 leads to a different eigendecomposition with identical eigenfunctions. Therefore, combining all these eigendecompositions, eigenvalue degeneracy only occurs if two eigenfunctions share the same eigenvalue for all $z_2 \in \mathcal{Z}_2$, which means that $f_{Z_2|\lambda,X}(z_2;v,x) = f_{Z_2|\lambda,X}(z_2;w,x)$ for all $z_2 \in \mathcal{Z}_2$. Recall that $Z_2 = Y_{R+1} \in \mathbb{R}$, while $\lambda \in \mathbb{R}^R$. Given the structure of the model, we get $f_{Z_2|\lambda,X}(z_2;v,x) = f_{Z_2|\lambda,X}(z_2;w,x)$ for all $z_2 \in \mathcal{Z}_2$ if $v'F_{R+1} = w'F_{R+1}$, which is clearly possible if $R > 1$. Hu and Schennach (2008) rule out this situation in their Assumption 4, but an analog of this assumption does not hold here if $R > 1$. Hence, compared to Hu and Schennach (2008), additional arguments are needed to solve the eigenvalue degeneracy problem. To do so, notice that, similar as in the linear model, we can rotate the outcomes in Z_1 and Z_2. Specifically, let $K \equiv \{k_1, k_2, \ldots, k_R\}$ be a set of any R integers between 1 and $R+1$ with $k_1 < k_2 < \ldots < k_R$ and let $k_{R+1} = \{1, \ldots, R+1\} \setminus K$. Define $Z_1^K \equiv (Y_{k_1}, \ldots, Y_{k_R})$ and $Z_2^K = Y_{k_{R+1}}$. For example, if $R = 2$ and $T = 5$, then we could take $K = \{2,3\}$ and $k_{R+1} = 1$ and thus, $Z_1^K = (Y_2, Y_3)$ and $Z_2^K = Y_1$. Let $\mathcal{Z}_1^K$ be the support of Z_1^K and, analogously to before, define the operators

$$L_{1,2,3}^K : \mathcal{L}_{bnd}^R(\mathcal{Z}_1^K) \to \mathcal{L}_{bnd}^R \quad \left(L_{1,2,3}^K m\right)(z_2,z_3) \equiv \int f_{Z_1^K, Z_2^K, Z_3|X}(z_1,z_2,z_3;x) m(z_1) dz_1$$

$$L_{1,3}^K : \mathcal{L}_{bnd}^R(\mathcal{Z}_1) \to \mathcal{L}_{bnd}^R \quad \left(L_{1,3}^K m\right)(z_3) \equiv \int f_{Z_1^K, Z_3|X}(z_1,z_3;x) m(z_1) dz_1$$

$$D_{2,\lambda}^K : \mathcal{L}_{bnd}^R(\Lambda) \to \mathcal{L}_{bnd}^R(\Lambda) \quad \left(D_{2,\lambda}^K m\right)(z_2,v) \equiv f_{Z_2^K|\lambda,X}(z_2;v,x) m(v).$$

Then using identical arguments to before, it can be shown that for all sets K

$$L_{1,2,3}^K (L_{1,3}^K)^{-1} = L_{3,\lambda} D_{2,\lambda}^K L_{3,\lambda}^{-1}.$$

It follows that $L_{1,2,3}^K (L_{1,3}^K)^{-1}$ has the same eigenfunctions for all K. Hence, by considering the eigendecomposition for all K, the eigenvalue degeneracy issue now only occurs if two or more eigenfunctions share the same eigenvalue for all operators, which is a similar idea to varying z_2 above. In terms of Y_t, this means that eigenvalue degeneracy arises if for $v \neq w$ it holds that $f_{Y_t|\lambda}(y_t;v) = f_{Y_t|\lambda}(y_t;w)$ for all $y_t \in \mathcal{Y}_t$ and all $t = 1, \ldots, R+1$. However, Assumptions N3(i), N4, and N6 imply that $M[Y_t \mid \lambda = v] = g_t(v'F_t)$, that g_t are strictly increasing functions, and that the matrix $(F_1 \ldots F_R)$ has full rank. Hence $f_{Y_t|\lambda}(y_t;v) = f_{Y_t|\lambda}(y_t;w)$ for all $y_t \in \mathcal{Y}_t$ and all $t = 1, \ldots, R+1$ implies that $v'(F_1 \ldots F_R) = w'(F_1 \ldots F_R)$, which in turn implies that $v = w$.

Third, I show that there is a unique *ordering* of the eigenfunctions which coincides with $L_{3,\lambda}$. Generally, the problem is that while the sets of eigenfunctions and eigenvalues are uniquely determined, these sets do not uniquely define the distribution of $\lambda \mid X$. In particular, let $\tilde{\lambda} = B(\lambda, x)$, where $B(\cdot,x)$ is a one-to-one transformation of λ, which may depend on x. Then $f_{Z_3|\lambda,X}(\cdot;v,x) = f_{Z_3|B(\lambda,x),X}(\cdot;B(v,x),x)$ and hence each eigenfunction could belong to $f_{Z_3|\tilde{\lambda},X}(\cdot;\tilde{v},x)$ for some $\tilde{v}$.[30] To solve the ordering issue, Hu and Schennach (2008) and Cunha *et al.* (2010) assume that there exists a known function Ψ such that $\Psi(f_{Z_3|\lambda,X}(\cdot;\lambda,x)) = \lambda$ (see Assumption 5 in Hu and Schennach (2008)). Notice that in the factor model discussed in this article, the assumptions already imply $M(Y_{R+1+r} \mid \lambda = v, X = x) = g_{R+1+r}(x, v_r)$. Hence, it might be tempting to impose the "normalization" $g_{R+1+r}(x, \lambda_r) = \lambda_r$ so that $M(Z_3 \mid \lambda, X = x) = \lambda$. However, as shown below, here the distribution of $Y \mid \lambda$ is identified without such an additional "normalization" of λ. Thus, imposing this "normalization" is only consistent with the model if g_{R+1+r} is linear in the second argument for all $r = 1, \ldots, R$, which is a strong assumption. Now to show that there is a unique ordering, first notice that both $\tilde{\lambda} = B(\lambda, x)$ and λ have to be consistent with the model. In particular, for $\tilde{\lambda}$ there has to exist strictly increasing and differentiable functions $\tilde{g}_t$ (with inverses $\tilde{h}_t$) such that

$$M(Y_{R+1+r} \mid \tilde{\lambda} = \tilde{v}, X = x) = \tilde{g}_{R+1+r}(x_{R+1+r}, \tilde{v}_r) \text{ for all } r = 1, \ldots, R.$$

In particular, the conditional median of Y_{R+1+r} only depends on the r'th element of $\tilde{v}_r$. Since

$$M(Y_{R+1+r} \mid \lambda = v, X = x) = M(Y_{R+1+r} \mid B(\lambda, x) = B(v, x), X = x)$$

it follows that $g_{R+1+r}(x_{R+1+r}, v_r) = \tilde{g}_{R+1+r}(x_{R+1+r}, B_r(v,x))$. Moreover, since $\tilde{g}_{R+1+r}$ is strictly increasing and differentiable, it has to hold that $B_r(\cdot,x)$ is differentiable. Since the left-hand side only depends on v_r, it follows that

30. To see why $B(\cdot,x)$ has to be one-to-one, notice that since the set of eigenfunctions is uniquely determined, for each v and w, there has to exist $B(v,x)$ and $B(w,x)$ such that $f_{Z_3|\lambda,X}(\cdot;v,x) = f_{Z_3|\tilde{\lambda},X}(\cdot;B(v,x),x)$ and $f_{Z_3|\lambda,X}(\cdot;w,x) = f_{Z_3|\tilde{\lambda},X}(\cdot;B(w,x),x)$. But as shown above, if $v \neq w$, then $f_{Z_3|\lambda,X}(\cdot;v,x) \neq f_{Z_3|\lambda,X}(\cdot;w,x)$ which immediately implies that $f_{Z_3|\tilde{\lambda},X}(\cdot;B(v,x),x) \neq f_{Z_3|\tilde{\lambda},X}(\cdot;B(w,x),x)$, and thus $B(v,x) \neq B(w,x)$.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

$\partial B_r(v,x)/\partial v_s = 0$ for all $s \neq r$. Hence, $B_r(v,x)$ only depends on v_r. Next, it also holds by independence of U_{R+1+r} and λ, conditional on X, that

$$P(Y_{R+1+r} \leq y \,|\, X=x, \lambda=v) = F_{U_{R+1+r}|X}(h_{R+1+r}(y,x_{R+1+r}) - v_r; x),$$

and therefore it has to hold that for some $\tilde{F}_{U_t|X}$

$$F_{U_{R+1+r}|X}(h_{R+1+r}(y,x_{R+1+r}) - v_r; x) = \tilde{F}_{U_{R+1+r}|X}(\tilde{h}_{R+1+r}(y,x_{R+1+r}) - B_r(v_r,x); x).$$

Then taking the ratio of the derivatives with respect to v_r and y yields

$$\frac{\tilde{h}'_{R+1+r}(y,x_{R+1+r})}{h'_{R+1+r}(y,x_{R+1+r})} = B'_r(v_r,x).$$

But since at $\bar{y}_{R+1+r}$ (recall that $x_t = \bar{x}_t$ for $t = R+2,\ldots,2R+1$), we get

$$\tilde{h}'_{R+1+r}(\bar{y}_{R+1+r}, \bar{x}_{R+1+r}) = h'_{R+1+r}(\bar{y}_{R+1+r}, \bar{x}_{R+1+r}) = 1,$$

it has to hold that $B_r(v_r,x) = v_r + d_r(x)$ for some functions $d_r(x)$. Moreover, for all $r = 1,\ldots,R$ it has to hold that $g_{R+1+r}(x_{R+1+r}, v_r) = \tilde{g}_{R+1+r}(x_{R+1+r}, v_r + d_r(x))$, or alternatively $\tilde{h}_{R+1+r}(y_{R+1+r}, x_{R+1+r}) = h_{R+1+r}(y_{R+1+r}, x_{R+1+r}) + d_r(x)$, where $y_{R+1+r} \equiv g_{R+1+r}(x_{R+1+r}, v_r)$. But since at $\bar{y}_{R+1+r}$ we have $\tilde{h}_{R+1+r}(\bar{y}_{R+1+r}, \bar{x}_{R+1+r}) = h_{R+1+r}(\bar{y}_{R+1+r}, \bar{x}_{R+1+r}) = 0$, it has to hold that $d_r(x) = 0$. Therefore, only $B(\lambda,x) = \lambda$ is consistent with the model.[31]

Since none of the three non-unique features can occur due to the assumptions and structure of the model, $L_{3,\lambda}$ and $D_{2,\lambda}$ are identified. By the relation $L_{3,\lambda}^{-1} L_{1,3} = L_{\lambda,1}$ it also holds that $L_{\lambda,1}$ is identified. The operator being identified is the same as the kernel being identified. Hence, $f_{Y,\lambda|X}(y,v;x)$ is identified for all $y \in \mathbb{R}^T$, $v \in \Lambda$, and $x \in \tilde{\mathcal{X}}$. Since λ_r has support on $\mathbb{R}$ for all $r = 1,\ldots,R$, g_{R+1+r} is identified for all $r = 1,\ldots,R$ because $M[Y_{R+1+r} \,|\, \lambda = v, X = x] = g_{R+1+r}(x_{R+1+r}, v_r)$ and $f_{Y,\lambda|X}$ is identified. Similarly $M[Y_t \,|\, \lambda = v, X = x] = g_t\left(x_t, v' F_t\right)$ for all $t < R+2$. If $R = 1$, then g_t is identified up to scale, which is fixed by Assumption N3. If $R > 1$, taking ratios of derivatives with respect to different elements of λ identifies $\frac{F_{tr}}{F_{ts}}$ for all $r,s = 1,\ldots,R$. Hence, again g_t is identified up to scale which is fixed. Therefore, g_t and F_t are identified.

Finally suppose that $f_X(x) > 0$ for all $x \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_T$. Then the previous arguments imply that g_t is identified for all $x_t \in \mathcal{X}_t$ and $t < R+2$. Next take any $x \in \mathcal{X}$. Since F_t is identified and g_t is identified for all $x_t \in \mathcal{X}_t$ and $t < R+2$, the arguments above imply that $f_{Y,\lambda|X}(y,v;x)$ is then identified for all $y \in \mathcal{Y}$, $v \in \Lambda$ by switching the roles of $(Y_1,\ldots,Y_R)$ and $(Y_{R+2},\ldots,Y_{2R+2})$ in the proof. Consequently, g_t and the distribution of $(U,\lambda) \,|\, X = x$ are identified for all $x \in \mathcal{X}$. ‖

A.3. *Proof of Theorem 2*

First fix $\bar{x} \in \mathcal{X}$ and $\bar{y}$ on the support of $Y \,|\, X = \bar{x}$ and define $d_t = \frac{\partial h_t(\bar{y}_t, \bar{x}_t)}{\partial y}$ for all t. Next let $F^3 = (F_{R+1} \cdots F_{2R+1})$, $\bar{F} = (F^3)^{-1} F$, $\tilde{F}_t = \bar{F}_t / d_t$ if $t = 1,\ldots,R+1$ and $\tilde{F}_t = \bar{F}_t$ if $t = R+2,\ldots,2R+1$. Let $\tilde{\lambda}' = (\lambda'(F^3)^{-1} - b')$ and $\tilde{\lambda}_r = \lambda_r / d_{R+1+r}$ for $r = 1,\ldots,R$, where b is chosen such that $b'\bar{F}_t + c_t = h_t(\bar{y}_t, \bar{x}_t)$ for $t = R+2,\ldots,2R+1$. Finally let $\tilde{h}_t(y,x) = \frac{h_t(y,x) - b'\bar{F}_t - c_t}{d_t}$ and $\tilde{U}_t = \frac{U_t - c_t}{d_t}$. Then we get $\tilde{h}_t(Y_t, X_t) = \tilde{\lambda}'\tilde{F}_t + \tilde{U}_t$, $\tilde{h}_t(\bar{y}_t, \bar{x}_t) = 0$ for all $t = R+2,\ldots,2R+1$, $\frac{\partial \tilde{h}_t(\bar{y}_t, \bar{x}_t)}{\partial y} = 1$ for all $t = 1,\ldots,T$, $\tilde{F}^3 = I_{R\times R}$, and $M[\tilde{U}_t \,|\, X, \tilde{\lambda}] = 0$. By Theorem 1, $\tilde{h}_t(\cdot, x_t)$, $\tilde{F}_t$ and the distribution of $\tilde{U}, \tilde{\lambda} \,|\, X = x$ are identified for all $x \in \mathcal{X}$. Thus, the distribution of $\tilde{C}_t = \tilde{\lambda}'\tilde{F}_t$ and $\tilde{g}_t\left(\tilde{x}_t, Q_{\alpha_1}[\tilde{C}_t \,|\, X = x] + Q_{\alpha_2}[\tilde{U}_t \,|\, X = x]\right)$ are identified for each t, all $\tilde{x} \in \mathcal{X}_t$, and $x \in \mathcal{X}$. Finally, it holds that

$$\begin{aligned}
&\tilde{g}_t\left(\tilde{x}_t, Q_{\alpha_1}[\tilde{C}_t \,|\, X = x] + Q_{\alpha_2}[\tilde{U}_t \,|\, X = x]\right) \\
&\qquad = g_t\left(\tilde{x}_t, \left(Q_{\alpha_1}[\tilde{C}_t \,|\, X = x] + Q_{\alpha_2}[\tilde{U}_t \,|\, X = x]\right) d_t + b'\bar{F}_t + c_t\right) \\
&\qquad = g_t\left(\tilde{x}_t, \left(\frac{Q_{\alpha_1}[C_t \,|\, X = x] - b'\bar{F}_t}{d_t} + \frac{Q_{\alpha_2}[U_t \,|\, X = x] - c_t}{d_t}\right) d_t + b'\bar{F}_t + c_t\right) \\
&\qquad = g_t\left(\tilde{x}_t, Q_{\alpha_1}[C_t \,|\, X = x] + Q_{\alpha_2}[U_t \,|\, X = x]\right).
\end{aligned}$$

31. When $R > 1$, it can be shown that $B(\lambda,x) = \lambda$ using only median independence and not full independence. Hence, even without independence of U_t and λ, imposing a "normalization" of the form $\Psi(f_{Z_3|\lambda}(\cdot;\lambda)) = \lambda$ is not without loss of generality.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

Similarly, since $P\left(\tilde{C}_t+\tilde{U}_t<e\,|\,X=x\right)=P\left(C_t+U_t<ed_t+b'\bar{F}_t+c_t\,|\,X=x\right)$ it follows that

$$\begin{aligned}\int \tilde{g}_t(\tilde{x}_t,e)\,dF_{\tilde{C}_t+\tilde{U}_t|X}(e;x) &= \int \tilde{g}_t(\tilde{x}_t,e)\,dF_{C_t+U_t|X}\left(ed_t+b'\bar{F}_t+c_t;x\right)\\ &= \int g_t\left(\tilde{x}_t,\left(\frac{e-b'\bar{F}_t-c_t}{d_t}\right)d_t+b'\bar{F}_t+c_t\right)dF_{C_t+U_t|X}(e;x)\\ &= \int g_t(\tilde{x}_t,e)\,dF_{C_t+U_t|X}(e;x)\end{aligned}$$

Analogous arguments yields $\tilde{g}_t(\tilde{x}_t,Q_\alpha[\tilde{C}_t+\tilde{U}_t\,|\,X=x])=g_t(\tilde{x}_t,Q_\alpha[C_t+U_t\,|\,X=x])$ and identification of $g_t(\tilde{x}_t,Q_\alpha[C_t+U_t\,|\,X])$ as well as identification of the unconditional quantities. ‖

Acknowledgments. This paper is a revised version of my job market paper. I am very grateful to Joel Horowitz as well as Ivan Canay and Elie Tamer for their excellent advice, constant support, and many helpful comments and discussions. I thank Stéphane Bonhomme and four anonymous referees for valuable suggestions, which helped to substantially improve the paper. I have also received helpful comments from James Heckman, Matt Masten, Konrad Menzel, Jack Porter, Diane Schanzenbach, Arek Szydlowski, Alex Torgovitsky, and seminar particatipants at various institutions. I thank Jan Bietenbeck for sharing his data and STATA code and for many helpful discussions. Financial support from the Robert Eisner Memorial Fellowship is gratefully acknowledged.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

REFERENCES

ACKERBERG, D., CHEN, X. and HAHN, J. (2012), "A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators", *The Review of Economics and Statistics*, **94**, 481–498.

AHN, S., LEE, Y. and SCHMIDT, P. (2001), "GMM Estimation of Linear Panel Data Models with Time-varying Individual Effects", *Journal of Econometrics*, **101**, 219–255.

—— (2013), "Panel Data Models with Multiple Time-varying Individual Effects", *Journal of Econometrics*, **174**, 1–14.

AI, C. and CHEN, X. (2003), "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions", *Econometrica*, **71**, 1795–1843.

ALTONJI, J. and MATZKIN, R. (2005), "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors", *Econometrica*, **73**, 1053–1102.

ANDREWS, D. (2005), "Cross-section Regression with Common Shocks", *Econometrica*, **73**, 1551–1585.

ARELLANO, M. and BONHOMME, S. (2012), "Identifying Distributional Characteristics in Random Coefficients Panel Data Models", *Review of Economic Studies*, **79**, 987–1020.

BAI, J. (2003), "Factor Models of Large Dimensions", *Econometrica*, **71**, 135–171.

—— (2009), "Panel Data Models with Interactive Fixed Effects", *Econometrica*, **77**, 1229–1279.

—— (2013), "Fixed-Effects Dynamic Panel Models, A Factor Analytical Method", *Econometrica*, **81**, 285–314.

BAI, J. and NG, S. (2002), "Determining the Number of Factors in Approximate Factor Models", *Econometrica*, **70**, 191–221.

BESTER, A. and HANSEN, C. (2009), "Identification of Marginal Effects in a Nonparametric Correlated Random Effects Model", *Journal of Business and Economic Statistics*, **27**, 235–250.

BIETENBECK, J. (2014), "Teaching Practices and Cognitive Skills", *Labour Economics*, **20**, 143–153.

Blundell, R. W., and J. L. Powell (2003), "Endogeneity in Nonparametric and Semiparametric Regression Models", in Dewatripont, M., Hansen, L. P. and Turnovsky, S. J., (eds), *Advances in Economics and Econonometrics: Theory and Applications, Eighth World Congress*, vol. 2. (Cambridge, UK: Cambridge University Press).

BONHOMME, S. and ROBIN, J.-M. (2008), "Consistent Noisy Independent Component Analysis", *Journal of Econometrics*, **149**, 12–25.

CARNEIRO, P., HANSEN, K. T. and HECKMAN, J. J. (2003), "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice", *International Economic Review*, **44**, 361–422.

CARROLL, R. J., CHEN, X. and HU, Y. (2010), "Identification and Estimation of Nonlinear Models using Two Samples with Nonclassical Measurement Errors", *Journal of Nonparametric Statistics*, **22**, 379–399.

CHAMBERLAIN, G. (1992), "Efficiency Bounds for Semiparametric Regression", *Econometrica*, **60**, 567–596.

CHEN, X. (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models", in Heckman, J. and Leamer, E., (eds), *Handbook of Econometrics*, Vol. 6 of *Handbook of Econometrics*, chap. 76. (Amsterdam, North-Holland: Elsevier) 5550–5623.

CHEN, X., TAMER, E. and TORGOVITSKY, A. (2011), "Sensitivity Analysis in Semiparametric Likelihood Models" (Working paper).

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

CHERNOZHUKOV, V., FERNANDEZ-VAL, I., HAHN, J. and NEWEY, W. (2013), "Average and Quantile Effects in Nonseparable Panel Models", *Econometrica*, **81**, 535–580.
CLOTFELTER, C. T., LADD, H. F. and VIGDOR, J. L. (2010), "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects", *Journal of Human Resources*, **45**, 655–681.
CUNHA, F. and HECKMAN, J. J. (2008), "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation", *Journal of Human Resources*, **43**, 738–782.
CUNHA, F., HECKMAN, J. J. and SCHENNACH, S. M. (2010), "Estimating the Technology of Cognitive and Noncognitive Skill Formation", *Econometrica*, **78**, 883–931.
DEE, T. S. (2007), "Teachers and the Gender Gaps in Student Achievement", *Journal of Human Resources*, **42**, 528–554.
DELAIGLE, A., HALL, P. and MEISTER, A. (2008), "On Deconvolution with Repeated Measurements", *The Annals of Statistics*, **36**, 665–685.
D'HAULTFOEUILLE, X. (2011), "On The Completeness Condition In Nonparametric Instrumental Problems", *Econometric Theory*, **27**, 460–471.
EVDOKIMOV, K. (2010), "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity" (Working paper).
EVDOKIMOV, K. and WHITE, H. (2012), "An Extension of a Lemma of Kotlarski", *Econometric Theory*, **28**, 925–932.
—— (2011), "Nonparametric Identification of a Nonlinear Panel Model with Application to Duration Analysis with Multiple Spells" (Working paper).
FAMA, E. F. and FRENCH, K. R. (2008), "Dissecting Anomalies", *Journal of Finance*, **63**, 1653–1678.
FAN, J. (1991), "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems", *The Annals of Statistics*, **19**, 1257–1272.
GALLANT, A. R. and NYCHKA, D. W. (1987), "Semi-nonparametric Maximum Likelihood Estimation", *Econometrica*, **55**, 363–390.
GRAHAM, B. and POWELL, J. (2012), "Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models", *Econometrica*, **80**, 2105–2152.
HECKMAN, J. J. and SCHEINKMAN, J. A. (1987), "The Importance of Bundling in a Gorman-Lancaster Model of Earnings", *The Review of Economic Studies*, **54**, 243–255.
HECKMAN, J. J., STIXRUD, J. and URZUA, S. (2006), "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior", *Journal of Labor Economics*, **24**, 411–482.
Hidalgo-CABRILLANA, A. and LOPEZ-MAYANY, C. (2015), "Teaching Styles and Achievement: student and Teacher Perspectives" (Working paper).
HODERLEIN, S. and WHITE, H. (2012), "Nonparametric Identification in Nonseparable Panel Data Models with Generalized Fixed Effects", *Journal of Econometrics*, **168**, 300–314.
HOLTZ-EAKIN, D., NEWEY, W. and ROSEN, H. S. (1988), "Estimating Vector Autoregressions with Panel Data", *Econometrica*, **56**, 1371–1395.
Hu, Y. (2008), "Identification and Estimation of Nonlinear Models with Misclassification Error using Instrumental Variables: A General Solution", *Journal of Econometrics*, **144**, 27–61.
Y. Hu and SCHENNACH, S. M. (2008), "Instrumental Variable Treatment of Nonclassical Measurement Error Models", *Econometrica*, **76**, 195–216.
HUANG, X. (2013), "Nonparametric Estimation in Large Panels with Cross Sectional Dependence", *Econometric Reviews*, **32**, 754–777.
IMBENS, G. W. and NEWEY, W. K. (2009), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity", *Econometrica*, **77**, 1481–1512.
JUDD, K. (1998), *Numerical Methods in Economics* (Cambridge, Massachusetts, USA: The MIT Press).
LAVY, V. (2016), "What Makes an Effective Teacher? Quasi-Experimental Evidence", *CESifo Economic Studies*, **62**, 88–125.
MADANSKY, A. (1964), "Instrumental Variables in Factor Analysis", *Psychometrika*, **29**, 105–113.
MATZKIN, R. (2003), "Nonparametric Estimation of Nonadditive Random Functions", *Econometrica*, **71**, 1339–1375.
MOON, H. R. and WEIDNER, M. (2015), "Linear Regression for Panel with Unknown Number of Factors as Interactive Fixed Effects", *Econometrica*, **83**, 1543–1579.
NEWEY, W. and POWELL, J. (2003), "Instrumental Variable Estimation of Nonparametric models", *Econometrica*, **71**, 1565–1578.
NEWEY, W. K. and McFADDEN, D. (1994), "Large Sample Estimation and Hypothesis Testing", in Engle, R. F. and McFadden, D., (eds), *Handbook of Econometrics*, vol. 4 of *Handbook of Econometrics*, chap. 36 (Amsterdam, North-Holland: Elsevier) 2111–2245.
PESARAN, M. H. (2006), "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure", *Econometrica*, **74**, 967–1012.
SCHWERDT, G. and WUPPERMANN, A. C. (2011), "Is Traditional Teaching Really all that Bad? A Within-Student Between-Subject Approach", *Economics of Education Review*, **30**, 365–379.
SHIU, J.-L. and HU, Y. (2013), "Identification and Estimation of Nonlinear Dynamic Panel Data Models with Unobserved Covariates", *Journal of Econometrics*, **175**, 116–131.
SU, L. and JIN, S. (2012), "Sieve Estimation of Panel Data Models with Cross Section Dependence", *Journal of Econometrics*, 169(1), 34–47.
WAYNE, A. J. and YOUNGS, P. (2003), "Teacher Characteristics and Student Achievement Gains: A Review", *Review of Educational Research*, **73**, 89–122.

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018

WILHELM, D. (2015), "Identification and Estimation of Nonparametric Panel Data Regressions with Measurement Error" (Working paper).
WILLIAMS, B., HECKMAN, J. and SCHENNACH, S. (2010), "Nonparametric Factor Score Regression with an Application to the Technology of Skill Formation" (Working paper).
ZEMELMAN, S., DANIELS, H. and HYDE, A. (2012), *Best Practice: Bring Standards to Life in America's Classrooms* (Portsmouth, New Hampshire, USA: Heinemann).

Downloaded from https://academic.oup.com/restud/article-abstract/85/3/1824/4105005
by University of Wisconsin-Madison Libraries user
on 25 June 2018