Estimation and Inference for Linear Models with Two-Way Unobserved Heterogeneity and Sparsely Matched Data

Valentin Verdier*

October 18, 2016

Abstract

I study the estimation of coefficients on finitely many observed covariates in linear models of panel data that include two terms of unobserved heterogeneity, each indexed by an observed indexing variable. I use an asymptotic framework where few observations are assigned to each value of either indexing variable. With this asymptotic framework, a two-way fixed effects regression does not necessarily preserve moment boundedness. I propose a simple alternative transformation that leads to an estimator that I show to be consistent and asymptotically normal when instrumental variables are strictly exogenous and in the presence of arbitrary dependence in the transitory shocks indexed by the same indexing variables that index unobserved heterogeneity. When instrumental variables are sequentially exogenous (e.g. dynamic models), I propose a new sequential transformation and estimator. Finally I provide estimates of the effect of class size reduction on student achievement using student-level data from North Carolina and a dynamic model of learning that accounts for sorting based on student unobserved ability and teacher and school unobserved quality.

Supplemental material: [https://goo.gl/r5jaLd](https://goo.gl/r5jaLd)

Keywords: Panel data, Two-way fixed effects, Dynamic models, Value-added models of student achievement; JEL codes: C23

---

*Assistant Professor, Department of Economics, University of North Carolina, Chapel Hill, NC 27599, United States. Tel.: +1 919-966-3962. E-mail address: vverdier@email.unc.edu.
1 Introduction

Models with two-way unobserved heterogeneity are frequently used in empirical work, either as an attempt to control for potentially confounding factors, or because features of the joint distribution of the unobserved heterogeneity terms are of interest. In the analysis of education data, student and teacher or school fixed effects are routinely used as a way to control for non-random selection on unobservables (e.g. Hanushek et al. (2003), Rivkin et al. (2005), Sund (2009), Clotfelter et al. (2010), Kingdon and Teal (2010), Rothstein (2010)). In the analysis of employer-employee matched data, models with worker and firm unobserved heterogeneity are often used to study features of the distribution of worker and firm productivity (e.g. Abowd et al. (1999), Postel-Vinay and Robin (2002), Card et al. (2013), Bonhomme et al. (2016)).

Here I expose linear models with two-way unobserved heterogeneity to an asymptotic framework where few observations are assigned to each value of either indexing variable. With education data for instance, this would correspond to having few time periods per student and few students per teacher, which seems relevant since most empirical applications follow students over only a few grades, and class sizes are limited by state laws or common practices. With employer-employee matched data, using an asymptotic framework where neither term of unobserved heterogeneity can be estimated precisely even after normalization can also be relevant depending on the size of firms, the number of observed time periods, and the strength of serial and cross-sectional dependence. A case where the asymptotic framework used here clearly does not apply is when there is an observation for each combination of values of both indexing variables, for instance international trade data where imports and exports for any pair of country are observed, or balanced panel data used with models that include individual and time effects only. In this case estimators can be defined after a simple three-way demeaning of the original data and complications only arise when using non-linear models, as studied for instance in Charbonneau (2014), Graham (2015), Fernández-Val and Weidner

1Models with two-way unobserved heterogeneity are not restricted to education or worker-firm matched data. For instance Chetty and Hendren (2016) and Finkelstein et al. (2016) consider households and patients assigned to geographical locations and models with two-way unobserved heterogeneity for income and measures of health.

2This is a known fact but is shown in Appendix A using the notation of this paper.
While estimation of features of the distribution of unobserved heterogeneity is of interest in many empirical studies, I leave the study of this question within the framework used in this paper to ongoing work and concentrate here on the estimation of coefficients on finite dimensional covariates for three reasons: These parameters are of primary interest in many applications, such as in the empirical application presented in this paper; It is necessary to estimate these parameters before estimating features of the distribution of unobserved heterogeneity; And estimating features of the distribution of unobserved heterogeneity necessitates stronger assumptions than the assumptions that are used in this paper.

With strictly exogenous covariates, a commonly used estimator for models with two-way unobserved heterogeneity relies on a two-way fixed effects regression, i.e. a regression on a set of indicators for each value of both indexing variables. I show that when using an asymptotic framework where few observations are assigned to each value of either indexing variable, the transformation obtained from a two-way fixed effects regression is not necessarily bounded in absolute row sum. Therefore a two-way fixed effects regression does not necessarily preserve moment boundedness of the original data. As a solution I propose a different transformation that applies two-way fixed effects regressions to many subsets of the data in which both indexing variables take only finitely many values. The resulting estimator is then shown to be consistent, asymptotically normal, and to have consistent standard errors in the presence of arbitrary dependence indexed by the same indexing variables that index unobserved heterogeneity. Models with endogenous covariates and strictly exogenous instrumental variables are easily accommodated by using an auxiliary model for exactly identifying instrumental variables.

With sequentially exogenous variables, i.e. when the set of exogenous instrumental vari-

\footnote{In particular several papers study how to compute such regressions, e.g. Abowd et al. (2002), Guimaraes and Portugal (2010), McAllister et al. (2012), Somani and Wolak (2015).}

\footnote{Cameron et al. (2011) discussed standard errors for estimators of linear models without unobserved heterogeneity but with arbitrary two-way dependence, taking asymptotic normality as given. The asymptotic results presented in this paper can be modified to fit the framework of Cameron et al. (2011) and show that it guarantees asymptotic normality of the estimator they consider. In addition, as will be discussed below, the standard errors of Cameron et al. (2011) can not be used here as they would not be consistent when estimating models with two-way unobserved heterogeneity.}
ables increases with time, as with dynamic models for instance, the estimator described above is not consistent for the parameters of interest in an asymptotic framework where few observations are assigned to each value of the indexing variables, much as the one-way fixed effects estimator is not consistent when instrumental variables are sequentially exogenous in a model with one-way unobserved heterogeneity and few observations per cross-sectional unit (Nickell (1981), Anderson and Hsiao (1981), Arellano and Bond (1991)). Unlike with one-way unobserved heterogeneity however, one cannot obtain moment conditions that exhaust the information for estimating the parameters of interest contained in models with two-way unobserved heterogeneity by considering each cross-sectional observation in isolation, as was done in Chamberlain (1992) for instance. Here I show that such moment conditions can be obtained for the model with two-way unobserved heterogeneity by a recursive argument and that these moment conditions lead to a relatively simple estimator. I show that the estimator that I define is consistent, asymptotically normal, and has consistent standard errors under flexible patterns of cross-sectional dependence.

There are discussions in empirical work on potential biases that can appear with existing estimators of models with two-way unobserved heterogeneity when few observations are assigned to each value of the indexing variable (e.g. Abowd et al. (2004), Andrews et al. (2008, 2012)), but little econometric work that can be used to provide asymptotic justification for estimation and inference. Since additive unobserved heterogeneity can be rewritten as the product of a vector of indicator variables and a vector of coefficients, the results presented here are related to the analysis of linear models with high dimensional control covariates, as in Huber (1973) or Cattaneo et al. (2016). Here however we need to propose new estimation methods and derive new asymptotic results because accommodating cross-sectional dependence and sequentially exogenous instruments will be important for the empirical application and because the two-way fixed effects regression is shown to obtain transformed variables that do not have bounded higher moments even though the original variables do, all of which invalidates the assumptions in Cattaneo et al. (2016). Jochmans and Weidner (2016) have obtained finite sample results for the estimators of the terms of unobserved heterogeneity in
linear models with two-way unobserved heterogeneity. They derive bounds for the variance of the estimator of each term of unobserved heterogeneity that depend on the inverse of the minimum positive eigenvalue of the Laplacian of the graph formed by each value of both indexing variables as vertices and the number of observations for each combination of those values as edges.\(^5\) In the framework used here, this minimum positive eigenvalue will generally converge to zero as sample size increases, so that the bounds given in Jochmans and Weidner (2016) increase unboundedly as sample size increases. The objective here is to show that quantities of interest can be estimated consistently and that asymptotically valid inference can be conducted even in a framework where neither term of unobserved heterogeneity can be estimated precisely.

In Section 2 I present the model and object of interest. In Section 3 I discuss the behavior of two-way fixed effects regressions with sparsely matched data. In Section 4 I define an estimator for models with strictly exogenous instruments and discuss its asymptotic properties. In Section 5 I discuss models with sequentially exogenous instruments. In Section 6 I present results from a brief Monte Carlo simulation study that exemplify the biases that originate from ignoring the presence of unobserved heterogeneity, ignoring the sequential exogeneity of instrumental variables, or treating one of the terms of unobserved heterogeneity as a finite dimensional set of parameters that can be estimated consistently. In Section 7 I use longitudinal data from North Carolina to estimate the effect of class size reductions on student achievement among elementary school students using a dynamic model of learning with student unobserved ability and teacher unobserved quality.

\(^5\)Here different results from graph theory, or rather Markov chain theory, applied to the same graph are used to show that a two-way fixed effects regression does not necessarily preserve moment boundedness when few observations are assigned to each value of the indexing variables.
2 Model and Object of Interest

I consider models for cross-sectional observations \( i \in \mathcal{N}, \mathcal{N} = \{1, \ldots, n\} \), and time periods \( t \in \mathcal{T}, \mathcal{T} = \{1, \ldots, T\} \), that can be written as:

\[
y_{it} = x_{it} \beta_0 + c_i + e_{dit} + u_{it} \tag{2.1}
\]

\[
E(u_{it}|Z_t, D) = 0 \tag{2.2}
\]

where \( y_{it} \) is an observed outcome variable, \( x_{it} \) are observed covariates, \( c_i \) captures unobserved time constant factors that are specific to a cross-sectional observation, \( d_{it} \in \mathcal{D}_n \), \( \mathcal{D}_n = \{1, \ldots, N_n\} \), is observed and indexes unobserved factors, \( e_{dit} \), that are common to all observations sharing the same value for \( d_{it} \), \( u_{it} \) are all other unobserved factors, \( Z_t = \{z_{it}\}_{i \in \mathcal{N}} \) collects instrumental variables \( z_{it} \) for all cross-sectional observations, and \( D = \{d_{it}\}_{i \in \mathcal{N}, t \in \mathcal{T}} \) collects the assignment to values of the indexing variable \( d_{it} \) of all cross-sectional observations at all time periods.

Throughout the paper, all distributional statements will be made conditional on \( D \). In order to shorten notation, this conditioning will be suppressed, or alternatively \( D \) will be treated as deterministic. Therefore the model is rewritten:

\[
y_{it} = x_{it} \beta_0 + c_i + e_{dit} + u_{it} \tag{2.3}
\]

\[
E(u_{it}|Z_t) = 0 \tag{2.4}
\]

\( z_{it} \) is assumed to be an increasing sequence of sets of instrumental variables. If \( z_{it} = z_{is} \) for all \( t \) and \( s \) in \( \mathcal{T} \), instruments are said to be strictly exogenous. If \( z_{it} \subset z_{it+1} \), instruments are said to be sequentially exogenous.

As discussed in the introduction, the objective throughout the paper is the estimation of \( \beta_0 \).

An example of the models considered here is a value-added model of student learning with...
unobserved student ability and unobserved teacher quality, which will be used in the empirical application:

\[ y_{it} = \rho_0 y_{it-1} + w_{it} \theta_0 + c_i + e_{dit} + u_{it} \quad (2.5) \]

\[ E(u_{it}|Y^{t-1}, W) = 0 \quad (2.6) \]

where \( Y^{t-1} = \{y_{is}\}_{i \in N, s \in \{1, \ldots, t-1\}} \), \( W = \{w_{it}\}_{i \in N, t \in T} \). In this example, \( y_{it} \) is student \( i \)'s achievement in year \( t \), \( d_{it} \) corresponds to the identity of the teacher who taught student \( i \) in year \( t \), so that \( e_{dit} \) can be interpreted as the unobserved quality of student \( i \)'s teacher in year \( t \). \( c_i \) represents student \( i \)'s unobserved ability. \( u_{it} \) could capture an education policy of interest.

Note that (2.4) only restricts the dependence of \( u_{it} \) on the exogenous variables \( Z_t \), while the dependence of \( c_i \) and \( e_{dit} \) on any other variable is left unrestricted. For example in the model of student learning (2.5) and (2.6), the level of student ability and teacher quality can vary systematically with the level of policy treatment captured by \( w_{it} \), and there can exist arbitrary dependence in the unobserved ability of students who attend the same classroom as well as between the unobserved ability of students and the unobserved quality of their teacher.

3 Two-Way Fixed Effects Regression with Sparsely Matched Data

Because (2.3) and (2.4) do not impose any restriction on the dependence between either term of unobserved heterogeneity and the observed exogenous variables, a natural approach to the estimation of \( \beta_0 \) can start with transformations of the data that are orthogonal to the terms of unobserved heterogeneity, independently of the values they take.

A natural choice for such a transformation is the two-way fixed effects regression. Stacking the data over cross-sectional observations and time, we can define \( y_n = [y_{it}]_{i \in N, t \in T} \), \( x_n = [x_{it}]_{i \in N, t \in T} \), \( c_n = [c_i]_{i \in N} \), \( e_n = [e_{id}]_{d \in D_n} \), \( u_n = [u_{it}]_{i \in N, t \in T} \), \( g_{n,1} = [1[i = j]]_{i \in N, t \in T} \), \( g_{n,2} = [1[d_{it} = d]]_{d \in D_n} \), \( g_n = [g_{n,1}, g_{n,2}] \). With this new notation, in the case where \( Z \equiv Z_t = \)}
Given \( t, s \in T \) (i.e. with strictly exogenous instrumental variables), (2.3) and (2.4) can be rewritten:

\[
y_n = x_n\beta_0 + g_n \begin{bmatrix} c_n \\ e_n \end{bmatrix} + u_n \quad (3.1)
\]

\[
E(u_n|Z) = 0 \quad (3.2)
\]

Define \( P_{g_n} = g_n(g_n'g_n)^{-}g_n' \) and \( M_{g_n} = I_nT - P_{g_n} \), where \( A^{-} \) denotes a generalized inverse for a matrix \( A \) and \( I_k \) is the identity matrix with \( k \) rows.

Then (3.1) and (3.2) are equivalent to:

\[
E(M_{g_n}(y_n - x_n\beta_0)|Z) = 0 \quad (3.3)
\]

\[
E(\begin{bmatrix} c_n \\ e_n \end{bmatrix}|Z) = (g_n'g_n)^{-}g_n'E(y_n - x_n\beta_0|Z)
\]

\[+ (I_n+D_n - (g_n'g_n)^{-}g_n'g_n)\xi_n \quad (3.4)\]

where \( \xi_n \) is an unrestricted \( (n + N_n) \times 1 \) vector.

Since no restriction is imposed on \( E(\begin{bmatrix} c_n \\ e_n \end{bmatrix}|Z) \), all of the information relevant for estimating \( \beta_0 \) is contained in (3.3).

Let \( \#A \) denote the cardinality of a countable set \( A \). In this section I study the behavior of two-way fixed effects regressions with an asymptotic framework where \( T \) is fixed and the number of observations assigned to each value of \( d_{it} \) is small, i.e. \( \#\{i,t\} \in \mathcal{N} \times \mathcal{T} : d_{it} = d \} \leq C, \forall d \in D_n, \forall n \) for some constant \( C < \infty \), so that the number of possible values for \( d_{it} \), \( N_n \), grows at a rate that is proportional to the number of cross-sectional observations, \( n \).

For instance, the empirical application in this paper uses observations on three cohorts of around 55,000 students, each with two repeated observations (grades 4 and 5), and approx-
imately 12,000 teachers, who are matched with around 15 students on average in each year (2009, 2010, 2011, 2012). Consequently an asymptotic framework that treats each student and each teacher as appearing only a few times in the data seems appropriate here, particularly when one allows for arbitrary dependence over time at the student level and within classrooms.

Lemma 1 shows that when few observations are assigned to each value of either indexing variable, it is possible for $||M_{gn}||_{\infty}$ to grow unboundedly as $n$ grows, where $||.||_{\infty}$ is the $L_{\infty}$ induced norm of a matrix, i.e. the maximum absolute row sum of a matrix.

Define $N_{n,d} = \# \{ \{ i,t \} \in \mathcal{N} \times \mathcal{T} : d_{it} = d \}$ and $N_{n,dd'} = \frac{1}{T} \sum_{i,s,t} 1[d_{it} = d, d_{is} = d']$. Define $D_n = \text{diag}(N_{n,d})$ and $P_n = D_n^{-1}[N_{n,dd'}]_{d',d} \in \mathcal{D}_n$. Then $P_n$ is a valid transition matrix for a Markov chain on $\mathcal{D}_n$, or equivalently the normalized adjacency matrix of the weighted undirected graph $\mathcal{G}_n = \{ \mathcal{D}_n, [N_{n,dd'}]_{d',d} \in \mathcal{D}_n \}$, since $\sum_{d' \in \mathcal{D}_n} N_{n,dd'} = N_{n,d}$.

Define the stochastic process $\xi_\tau$ as a Markov chain on $\mathcal{D}_n$ with transition probability matrix $P_n$. Define $d'' H_{dd'} = P_d(\xi_\tau \text{ reaches } d' \text{ before } d'')$, where $P_d(.)$ denotes the probability conditional on being in the initial state $d$, if there is a path from $d$ to $d'$ or $d''$ in $\mathcal{G}_n$, and $d'' H_{dd'} = 0$ otherwise.

Lemma 1 shows that this quantity, which is an artifact of the structure of $\mathcal{G}_n$, can be used to determine the magnitude of the absolute row sums of the two-way fixed effects regression transformation and to show that it can grow unboundedly under the asymptotic framework used here.

**Lemma 1.** Define:

$$b_n = \max_{d_1,d_2 \in \mathcal{D}_n : N_{n,d_1,d_2} > 0} \sum_{d_3,d_4 \in \mathcal{D}_n : N_{n,d_3,d_4} > 0} |d_2 H_{d_3d_1} - d_2 H_{d_4d_1}| \quad (3.5)$$

---

8With one-way fixed effects or two-way fixed effects and exhaustive matching of all values of the two indexing variables, fixed effects regressions amount to simple demeanings as noted in the introduction and Appendix A, so that $||M_{gn}||_{\infty}$ is equal to a constant regardless of sample size.

9Note that the results of Lemma 1 could not be obtained by considering the inverse of the minimum positive eigenvalue of $I_{N_n} - P_n$ used in Jochmans and Weidner (2016) since there are assignments of $\{i,t\} \in \mathcal{N} \times \mathcal{T}$ to $d \in \mathcal{D}_n$ for which this quantity grows unboundedly as $n \to \infty$, $N_{n,d} \leq C < \infty \forall d \in \mathcal{D}_n \forall n$, but for which $||M_{gn}||_{\infty}$ remains finitely bounded.
Under $N_{n,d} \leq C < \infty \forall d \in D_n$, there are constants $C_1 < \infty$ and $c_2 > 0$ such that, $\forall n$:

$$||M_{g_n}||_\infty \leq C_1 b_n$$  \hspace{1cm} (3.6)

and for $T = 2$:

$$||M_{g_n}||_\infty \geq c_2 b_n$$  \hspace{1cm} (3.7)

There is a sequence of assignments of $\{i, t\} \in N \times T$ to $d \in D_n \forall n$, such that $N_{n,d} \leq C < \infty \forall d \in D_n, \forall n$, and as $n \to \infty$, $||M_{g_n}||_\infty \to \infty$.

The proof of Lemma 1 is given in Appendix B.

As a result, even though it exhausts the information from our model in finite samples, a two-way fixed effects regression does not necessarily preserve moment boundedness of the original data when using an asymptotic framework where $T$ and $N_{n,d}$ are finitely bounded. While it is possible that asymptotic results could still be obtained for estimators that make use of transformed data with unbounded higher moments, here I choose to use a different transformation than the two-way fixed effects regression which will lead to consistent and asymptotically normal estimators with consistent standard errors\footnote{For instance normalizing the data using the inverse of $||M_{g_n}||_\infty$ is not a clear solution because of the discrepancy between $||M_{g_n}||_\infty$ and $||M_{g_n}||_2 = 1$ which would lead to the analogue of Assumption 4-b below being violated when $M_{g_n}$ and a normalization is used instead of $M_n$ and when $u_{it}$ is homoscedastic and serially and cross-sectionally uncorrelated for instance.} although in general it will be responsible for some loss of information for estimating $\beta_0$ compared to the information for estimating $\beta_0$ contained in (3.1) and (3.2).

Define $S_n$ to be a collection of subsets of $N$ such that $\#\{d \in D_n : d_{it} = d, i \in \sigma, t \in T\} \leq C \forall \sigma \in S_n$ and $\#\{\sigma \in S_n : i \in \sigma\} \leq C \forall i \in N, \forall n$ for some constant $C < \infty$. Note that $S_n$ does not necessarily form a partition of $N$. For $\sigma \in S_n$, define $g^{(\sigma)}_n = [1[i = j]_{i \in \sigma, t \in T}, 1[d_{it} = d]]_{i \in D_n, t \in T}$ and $S_{n,\sigma} = [1[i = i', t = t']]_{i \in \sigma, t \in T}$. We will make use of the transformation:

$$M_n = [S_{n,\sigma}^{\prime}]^{\sigma \in S_n} \text{diag}([M_{g_n(\sigma)}]_{\sigma \in S_n})[S_{n,\sigma}]_{\sigma \in S_n}$$  \hspace{1cm} (3.8)
where \( \text{diag}\{A_r\}_{r=1,\ldots,R} = \begin{bmatrix}
A_1 \\
0 & A_2 \\
... & ... \\
0 & ... & 0 & A_R
\end{bmatrix}. \)

When \( N_{n,d} \leq C \forall d \in D_n, \# \{ d : d_{it} = d, i \in \sigma, t \in T \} \leq C \forall \sigma \in S_n, \) and \( \# \{ \sigma \in S_n : i \in \sigma \} \leq C \forall i \in N \), it can be shown that \( \| M_n \|_\infty \leq C' \forall n \) for some constant \( C' < \infty \) by using the results shown in the proof of Lemma 1.

In the empirical application of section 7, there will be a natural way of defining such groups since students and teachers are contained within schools and each school contains a relatively small number of teachers. The Monte Carlo simulation study of section 6 also proposes another natural way of defining \( S_n \).

We will then use for estimation of \( \beta_0 \) the moment conditions:

\[
E(M_n(y_n - x_n \beta_0)|Z) = 0 \tag{3.9}
\]

4 Estimation with Strict Exogeneity

In this section I study the estimation of \( \beta_0 \) when instrumental variables are strictly exogenous, i.e. \( z_{it} = z_{is} \forall t, s \in T \). Define \( Z = Z_t \) for \( t \in T \), so that here \( \text{(2.4)} \) is rewritten \( E(u_{it}|Z) = 0 \).

The estimator I study is a relatively simple extension of linear estimators that rely on two-way fixed effects regression. Define \( z_{it} = E(x_{it}|Z) \) and \( \tilde{z}_n = [z_{it}]_{t \in T, i \in N} \), and define the potentially unfeasible estimator:

\[
\hat{\beta}_n = (\tilde{z}_n' M_n x_n)^{-1} \tilde{z}_n' M_n y_n \tag{4.1}
\]

\( \hat{\beta}_n \) is potentially unfeasible because if \( x_{it} \) is not part of \( z_{it} \), the form of \( E(x_{it}|Z) \) is unknown. A feasible estimator can be obtained by replacing \( \tilde{z}_{it} \) by a parsimonious model for \( E(x_{it}|Z) \) that is estimated in a first stage. As in Verdier (2016), it is likely that one can leverage cross-sectional dependence in order to obtain stronger instruments, and hence a more precise estimator.
The rest of the section lists assumptions under which the estimator $\hat{\beta}_n$ defined by (4.1) is consistent, asymptotically normal, has consistent standard errors, and is asymptotically efficient.

The first assumption imposes restrictions on the form of cross-sectional dependence in the transitory shocks $u_{it}$.

**Assumption 1.** Consider any $A,B \subset \mathcal{N} \times \mathcal{T}$ such that $\forall \{i,t\} \in A, \{j,s\} \in B$ with $i = j$, or $t = s$ and $d_{it} = d_{js}$. Then:

$$F(\{u_{it}\}_{i,t} \in A, \{u_{it}\}_{i,t} \in B \mid Z) = F(\{u_{it}\}_{i,t} \in A \mid Z) F(\{u_{it}\}_{i,t} \in B \mid Z)$$  \hspace{1cm} (4.2)

Assumption 1 does not impose any restriction on the serial dependence within cross-sectional observations or on the contemporaneous cross-sectional dependence among observations indexed by the same value of $d_{it}$, but it imposes independence otherwise. Dependence at a level higher than $d_{it}$ or $i$ can easily be accommodated as long as $d_{it}$ or $i$ take finitely many values at the level at which dependence exists. In our empirical example, Assumption 1 allows for arbitrary dependence among shocks to a student’s learning over time and among shocks to all students who attend the same classroom, but imposes independence otherwise. The assumption of independence across classrooms for different students could be relaxed, for example, to independence across schools for different students.

The second assumption imposes restrictions on the moments of the data. Let $\lambda_{\text{min}}(.)$ denote the smallest eigenvalue of a matrix, $||.||^r = (||.||_r)^r$ where $||.||_r$ is the $L_r$ vector norm, and $\sigma(.)$ to be the sigma-algebra generated by a list of random variables.

**Assumption 2.** There is a constant $c > 0$ and a random variable $B \in \sigma(Z)$ with $E(B) < \infty$ such that:

a) $\lambda_{\text{min}}(\frac{1}{n} E(\tilde{z}_n' \tilde{M} \tilde{z}_n)) \geq c, \forall n \geq N.$

b) $\lambda_{\text{min}}(\frac{1}{n} \text{Var}(\tilde{z}_n' M_n u_n)) \geq c, \forall n \geq N.$

c) $E(|u_{it}|^{4+\delta} \mid Z) \leq B, \ E(||x_{it}||^{4+\delta} \mid Z) \leq B \ a.e.$

Assumption 2-a is a standard assumption of relevance of the instrumental variables. As-
sumption 2-b is also standard and requires asymptotic ignorability of a particular summand. Assumption 2-c bounds the conditional higher moments of $u_{it}$ and $x_{it}$. Note that Assumption 2-c is stated conditionally on $Z$. This will be used for proving asymptotic normality of the estimator in this section when Assumptions 1 and 2 are used jointly with Assumption 3 below. If Assumption 4 below is used instead of Assumption 3, Assumption 2-c can be relaxed to unconditional moment boundedness.

Finally we need to restrict the cross-sectional dependence that exists in the covariates and the instrumental variables. The first possibility corresponds to the case where cross-sectional observations belong to possibly large groups of observations that are independent.

**Assumption 3.** $\forall n$, there exists a partition $\mathcal{P}_n$ of $\mathcal{N}$ such that \{$\{d : d_{it}, i \in p, t \in T\}\}_{p \in \mathcal{P}_n}$ also forms a partition of $\mathcal{D}_n$, and $\forall p \in \mathcal{P}_n$ \(F(\{u_{it}\}_{i \in p, t \in T} | Z) = F(\{u_{it}\}_{i \in p, t \in T} | \{z_{it}\}_{i \in p, t \in T})\), and $\forall p', p \neq p'$, \(F(\{x_{it}, z_{it}\}_{i \in p, t \in T}, \{x_{it}, z_{it}\}_{i \in p', t \in T}) = F(\{x_{it}, z_{it}\}_{i \in p, t \in T}) F(\{x_{it}, z_{it}\}_{i \in p', t \in T})\) and $\# p, # p' \leq C$ for a constant $C < \infty$. $\# \mathcal{P}_n \to \infty$ as $n \to \infty$.

Assumption 3 allows for flexible patterns of cross-sectional dependence since it does not restrict the strength of the cross-sectional dependence that exists within each set $p \in \mathcal{P}_n$, and these sets can be large relative to $n$ as long as they are of similar size and the number of sets in $\mathcal{P}_n$ is large. Note that the rate of convergence of the estimator in Proposition 1 below only depends on $n$ and not on $\# \mathcal{P}_n$.

In the empirical application of Section 7, school districts would be a natural example of sets that form a partition of the data and could be argued to be independent. However there is some movement of students and teachers across school districts, so the assumption that $\mathcal{P}_n$ forms a partition of $\mathcal{N}$ and that \{$\{d : d_{it}, i \in p, t \in T\}\}_{p \in \mathcal{P}_n}$ forms a partition of $\mathcal{D}_n$ seems inappropriate.

Instead of Assumption 3, an alternative assumption can be used that imposes that the cross-sectional dependence in the covariates and the instrumental variables decays with some measure of distance between locations, which in the empirical application could be the location of a student’s home for instance.

**Assumption 4.** Observations $\{i, t\} \in \mathcal{N} \times \mathcal{T}$ are associated with locations in $\mathbb{R}^q$, $q \geq 1$,
denoted by \( l_{it} \). Define \( Q = \{ l_{it} : i \in \mathbb{N}, t \in \mathcal{T} \} \) to be the lattice formed by all possible locations, and define the metric \( \rho(l, l') = \max_{1 \leq r \leq q}[|l[r] - l'[r]|] \) for any \( l, l' \in Q \). We have:

a) \( \rho(l, l') \geq 1 \forall l, l' \in Q, l \neq l' \).

b) \( \{\{x_{it}, z_{it}, s_{it}\} : i \in \mathbb{N}, t \in \mathcal{T}, n \in \mathbb{N}\} \) is an \( \alpha \)-mixing random field on the lattice \( \tilde{Q} = \{\{i, l_{it}\} : i \in \mathbb{N}, t \in \mathcal{T}\} \) equipped with the semi-metric \( \tilde{\rho}(\{i, l\}, \{j, l'\}) = 1[i \neq j] \rho(l, l') \) for \( \{i, l\}, \{j, l'\} \in \tilde{Q} \) with \( \alpha \)-mixing coefficients satisfying Assumption 3 of Jenish and Prucha (2012).

c) \( \forall A \subset \mathbb{N} \times \mathcal{T}, \text{let } B = \{\{j, s\} : \exists \{i, t\} \in A \text{ with } i = j, \text{ or } t = s \text{ and } d_{it} = d_{js}\}, \text{ and } F(\{u_{it}\}_{i, t} \in A | Z) = F(\{u_{it}\}_{i, t} \in A | \{z_{it}\}_{i, t} \in B). \)

d) \( \max_{i, j \in \sigma, \sigma \in \mathcal{S}_n, t, s \in \mathcal{T}} \rho(l_{it}, l_{js}) \leq C \) and \( \max_{i, j \in \mathbb{N}, t \in \mathcal{T}, d_{it} = d_{jt}} \rho(l_{it}, l_{jt}) \leq C \forall n \) for some constant \( C < \infty \).

Assumption 4-b imposes that the strength of cross-sectional dependence decays as distance increases, where distance is given by a semi-metric\(^{11}\) instead of a metric in order to accommodate arbitrary serial correlation at the cross-sectional level. Since \( \tilde{\rho}(\{i, l_{it}\}, \{j, l_{js}\}) \not\leq \tilde{\rho}(\{i, l_{it}\}, \{k, l_{kr}\}) + \tilde{\rho}(\{k, l_{kr}\}, \{j, l_{js}\}) \) and \( \tilde{\rho}(\{i, l_{it}\}, \{k, l_{kr}\}) = 0, l_{it} \neq l_{kr} \), for only finitely many choices of \( \{k, r\} (k = i, r \in \mathcal{T}, r \neq t) \), proofs of asymptotic results with spatial dependence are easily extended to this choice of semi-metric compared to using a metric such as \( \rho \) because both (semi-)metrics lead to equivalent cardinalities for basic sets in \( Q \) or \( \tilde{Q} \), as shown in Appendix C.

Assumption 4-a) is standard and corresponds to an increasing domain asymptotic framework as opposed to an infill asymptotic framework. Assumption 4-c restricts the form of distributional heteroscedasticity in \( u_{it} \) so that it corresponds to the form of cross-sectional independence of Assumption 1. Assumption 4-d imposes that the sets in \( \mathcal{S}_n \) contain cross-sectional observations that are relatively close to each other with respect to the measure of distance used throughout Assumption 4 and that cross-sectional observations assigned to the same value of \( d_{it} \) in the same time period are relatively close to each other. This is a natural assumption for the empirical application of section 7 for instance, where \( s \in \mathcal{S}_n \) corresponds

---

\(^{11}\)\( \tilde{\rho} \) is positive and symmetric, but does not satisfy definiteness and the triangle inequality
to all students who attend a particular school or one of a small number of neighboring schools and where \(d_{it}\) in a given year corresponds to the classroom that a student attends.

Under the assumptions above, the estimator defined by (4.1) is consistent and asymptotically normal.

**Proposition 1.** Under (2.3) and (2.4), when \(z_{it} = z_{is} \quad \forall t, s \in T\), if Assumptions 1 and 2, and Assumption 3 or Assumption 4 hold, as \(n \to \infty\) while \(N_{n,d} \leq C \quad \forall d \in D_n\), \#\{d : d_{it} = d, i \in \sigma, t \in T\} \leq C \quad \forall \sigma \in S_n\), and \#\{\sigma \in S_n : i \in \sigma\} \leq C \quad \forall i \in N, C < \infty, \forall n:\n
\[
\hat{\beta}_n \overset{p}{\to} \beta_0
\]  

(4.3)

and:

\[
\sqrt{n}V_n^{-\frac{1}{2}}A_n(\hat{\beta}_n - \beta_0) \overset{d}{\to} N(0, I)
\]  

(4.4)

where \(A_n = \frac{1}{n} E(\tilde{z}_n^\prime M_n \tilde{u}_n)\) and \(V_n = \frac{1}{n} \text{Var}(\tilde{z}_n^\prime M_n \tilde{u}_n)\).

Consistent standard errors are also available for the estimator defined in this section.

Let \(\hat{u}_n = y_n - x_n \hat{\beta}_n\). Let \(\delta_t(\sigma) = \{d \in D_n : \exists i \in \sigma, d_{it} = d\}\) for any \(\sigma \in S_n\) and \(t \in T\). Let \(S_n = [1\exists \sigma, \sigma' \in S_n, t \in T : i \in \sigma, j \in \sigma', \sigma \cap \sigma' \neq \emptyset \text{ or } \delta_t(\sigma) \cap \delta_t(\sigma') \neq \emptyset]_{i \in N, t \in T}\). Let \((A \odot B)\) denote the element-by-element (Hadamard) product of two matrices \(A\) and \(B\) of equal dimensions. Let \(||.||\) denote any matrix norm.

**Proposition 2.** Under (2.3) and (2.4), when \(z_{it} = z_{is} \quad \forall t, s \in T\), if Assumptions 1 and 2, and Assumption 3 or Assumption 4 hold, as \(n \to \infty\) while \(N_{n,d} \leq C \quad \forall d \in D_n\), \#\{d : d_{it} = d, i \in \sigma, t \in T\} \leq C \quad \forall \sigma \in S_n\), and \#\{\sigma \in S_n : i \in \sigma\} \leq C \quad \forall i \in N, C < \infty, \forall n:\n
\[
||\frac{1}{n} \tilde{z}_n^\prime M_n x_n - A_n|| \overset{p}{\to} 0
\]  

(4.5)

\[
||\frac{1}{n} \tilde{z}_n^\prime (M_n \hat{u}_n \hat{u}_n^\prime M_n' \odot S_n) \tilde{z}_n - V_n|| \overset{p}{\to} 0
\]  

(4.6)

Note that, unlike when estimating models with one-way unobserved heterogeneity[12] clustered standard errors using \(i\) and \(d_{it}\) as indexing variables as in Cameron et al. (2011) would

---

not lead to consistent standard errors here since, defining \( S_n^2 = \{[1[i = j or d_{it} = d_{js}]])_{i,j \in N, s \in T} \),
\[ E(\tilde{z}_n' M_n u_n u_n' M_n' \tilde{z}_n) \neq E(\tilde{z}_n' (M_n u_n u_n' M_n' \odot S_n^2) \tilde{z}_n), \] and \( E(\tilde{z}_n u_n) = E(\tilde{z}_n u_n' \odot S_n^2) \) but \( M_n(u_n[c_n', e_n'] g_n' \odot S_n^2) M_n' \) are not equal to zero and not asymptotically ignorable.

The next proposition shows that the estimator defined in this section is asymptotically efficient in the class of instrumental variable estimators when transitory shocks are homoscedastic and serially and cross-sectionally uncorrelated and when \( S_n \) forms a partition of the data.

**Proposition 3.** If \( S_n \) is a partition of \( N \), \( \{\{d_{it}\}_{i \in \sigma, t \in T}\}_{\sigma \in S_n} \) is a partition of \( D_n \), and:

\[
\begin{align*}
\text{Cov}(u_{it}, u_{js}|Z) &= \begin{cases} 
\sigma_u^2 & \text{if } i = j, \ t = s \\
0 & \text{otherwise}
\end{cases} 
\tag{4.7}
\end{align*}
\]

then:

\[
B_n^{-1} W_n B_n^{-1}' - A_n^{-1} V_n A_n^{-1}
\tag{4.8}
\]

is positive semi-definite \( \forall n \), where \( B_n = \frac{1}{n} E(\tilde{z}_n' M_n x_n) \) and \( W_n = \frac{1}{n} \text{Var}(\tilde{z}_n' M_n u_n) \) and \( \tilde{z}_n \) is any \( nT \times \text{dim}(\beta_0) \) function of \( Z \).

Note that considering the case where \( \tilde{z}_n \) is a matrix of exactly identifying instrumental variables is without loss of generality since optimal weighting can be applied to over-identifying instruments.

The proofs of Propositions 1-3 are found in Appendix C.

## 5 Estimation with Sequential Exogeneity

In this section I consider the case where instrumental variables are sequentially exogenous, i.e. \( Z_t \subset Z_s \) for \( s > t \). In the first part of the section I study how the two-way fixed effects regression should be modified for estimating models with sequentially exogenous instrumental variables. In the second part of the section I define a new estimator and list assumptions.
under which it is consistent, asymptotically normal, has consistent standard errors, and is asymptotically efficient.

5.1 Two-Way Fixed Effects Regression with Sequentially Exogenous Instruments

When \( Z_t \neq Z_s \) for \( t \neq s \), under (2.3) and (2.4), \( E(M_{g_n}u_n|Z_s) = 0 \) only for \( s = 1 \). Therefore in general the two-way fixed effects regression \( M_{g_n} \) does not preserve all of the information available in (2.3) and (2.4) for estimating \( \beta_0 \) when instrumental variables are sequentially exogenous. This is similar to models with one-way fixed effects, see e.g. Nickell (1981), Anderson and Hsiao (1981), Arellano and Bond (1991).

Define the forward filtering transformations \( M_{1,n,t} = \left( \frac{T-t}{T-t+1} \right)^{\frac{1}{2}} [1[i = i', t = t'] - \frac{1}{T-t+1}1[i = i', t' \geq t]] \in \mathcal{N}, t' \in \mathcal{T} \) for \( t \in \mathcal{T} \) and \( M_{1,n} = [M_{1,n,t}]_{t \in \mathcal{T}} \). Define \( \mathcal{g}_{2,n} = M_{1,n}\mathcal{g}_{2,n} \), and \( \mathcal{P}_{\mathcal{g}_{2,n}} = \mathcal{g}_{2,n}(\mathcal{g}_{2,n}^{'}\mathcal{g}_{2,n})^{-1} \mathcal{g}_{2,n}^{'} \).

Estimators for \( \beta_0 \) based on models with one-way fixed effects with sequentially exogenous instruments would treat \( e_n \) as a vector of parameters to estimate, and in general will not be consistent for \( \beta_0 \). For instance it is clear that when \( Z_t \neq Z_s \) for \( t \neq s \), (2.3) and (2.4) do not imply \( E(P_{\mathcal{g}_{2,n}}u_n|Z_t) = 0 \) for \( t > 1 \), and that when \( N_n \) grows unboundedly as \( n \) grows \( (\mathcal{g}_{2,n}^{'}\mathcal{g}_{2,n})^{-1} \mathcal{g}_{2,n}^{'}u_n \) is not asymptotically ignorable either.

Here I develop a transformation that leads to moment conditions that exhaust the information for estimating \( \beta_0 \) found in (2.3) and (2.4) when instrumental variables are sequentially exogenous.

Define \( y_{n,t} = [y_{it}]_{i \in \mathcal{N}}, x_{n,t} = [x_{it}]_{i \in \mathcal{N}}, \mathcal{g}_{2,n,t} = [1[d_{it} = d]]_{i \in \mathcal{N}} \).

(2.3) and (2.4) for \( t = T \) are equivalent to:

\[
E(c_n|Z_T) = E(y_{n,T} - x_{n,T}\beta_0 - \mathcal{g}_{2,n,T}e_n|Z_T)
\] (5.1)

Define \( y_{n,t} = M_{1,n,t}y_n, x_{n,t} = M_{1,n,t}x_n, \mathcal{g}_{2,n,t} = M_{1,n,t}\mathcal{g}_{2,n} \) for \( t = 1, ..., T - 1 \).

Since no restriction is imposed on \( E(c_n|Z_T) \), (5.1) contains no information for estimating \( \beta_0 \), and the information for estimating \( \beta_0 \) contained in (2.3) and (2.4) for \( t = 1, ..., T \) is
equivalent to the information contained in:

$$E(\dot{y}_{n,t} - \dot{x}_{n,t} \beta_0 - \dot{g}_{2,n,t} e_n | Z_t) = 0 \forall t = 1, ..., T - 1$$  \hfill (5.2)$$

With one-way fixed effects only, a similar result is found in Chamberlain (1992) for instance. With two-way fixed effects, we cannot use (5.2) for estimation directly since $e_n$ cannot be treated as a vector of parameters that can be estimated consistently.

For $t = T - 1$, (5.2) is equivalent to:

$$E(e_n | Z_{T-1}) = (\dot{g}_{2,n,T-1} \dot{y}_{n,T-1} - \dot{x}_{n,T-1} \beta_0 | Z_{T-1})$$

$$+ (I_{N_n} - (\dot{g}_{2,n,T-1} \phi_{2,n,T-1})^{-1} \dot{g}_{2,n,T-1} \phi_{2,n,T-1}) \xi_{T-1}$$  \hfill (5.3)$$

$$E(M_{g_{2,n,T-1}} (\dot{y}_{n,T-1} - \dot{x}_{n,T-1} \beta_0) | Z_{T-1}) = 0$$  \hfill (5.4)$$

where $\xi_{T-1}$ is an unrestricted $N_n \times 1$ vector.

Since $E(e_n | Z_{T-1})$ is unrestricted, no information for estimating $\beta_0$ is found in (5.3), and the information for estimating $\beta_0$ found in (5.2) for $t = T - 1$ is equivalent to the information found in (5.4).

Define $a_{n,T-1}(\beta_0) = (\phi_{2,n,T-1} \phi_{2,n,T-1})^{-1} \phi_{2,n,T-1} (\dot{y}_{n,T-1} - \dot{x}_{n,T-1} \beta_0)$ and $B_{n,T-1} = I_{N_n} - (\phi_{2,n,T-1} \phi_{2,n,T-1})^{-1} \phi_{2,n,T-1} \phi_{2,n,T-1}$.

The information for estimating $\beta_0$ found in (5.2) for $t = T - 1, T - 2$ is equivalent to the information contained in (5.4) and:

$$E(\dot{y}_{n,T-2} - \dot{x}_{n,T-2} \beta_0 - \dot{g}_{2,n,T-2} a_{n,T-1}(\beta_0) - \dot{g}_{2,n,T-2} B_{n,T-1} \xi_{T-1} | Z_{T-2}) = 0$$  \hfill (5.5)$$

Define $\ddot{g}_{2,n,T-2} = \dot{g}_{2,n,T-2} B_{n,T-1}$.
in turn is equivalent to:

\[
E(\xi_{T-1} | Z_{T-2}) = (\hat{g}_{2,n,T-2}^\prime \hat{g}_{2,n,T-2} - \hat{g}_{2,n,T-2})E(\dot{y}_{n,T-2} - \dot{x}_{n,T-2}\beta_0 - \dot{g}_{2,n,T-2}a_{n,T-1}(\beta_0)) | Z_{T-2}) \\
+ (I_n - (\hat{g}_{2,n,T-2}^\prime \hat{g}_{2,n,T-2} - \hat{g}_{2,n,T-2}^\prime \hat{g}_{2,n,T-2})\xi_{T-2})
\]

\[
E(M_{\hat{g}_{2,n,T-2}}(\dot{y}_{n,T-2} - \dot{x}_{n,T-2}\beta_0 - \dot{g}_{2,n,T-2}a_{n,T-1}(\beta_0)) | Z_{T-2}) = 0
\]

(5.6)

where \(\xi_{T-2}\) is an unrestricted \(N_n \times 1\) vector.

Since \(E(\xi_{T-1} | Z_{T-2})\) is unrestricted, we see that the information for estimating \(\beta_0\) found in (5.2) for \(t = T - 1, T - 2\) is equivalent to the information contained in (5.4) and (5.7).

Define:

\[
\ddot{g}_{2,n,T} = 0
\]

\[
B_{n,T} = I_n
\]

(5.8)

(5.9)

and, for \(t = T - 1, \ldots, 1\):

\[
\ddot{g}_{2,n,t} = \dot{g}_{2,n,t} B_{n,t+1}
\]

\[
B_{n,t} = B_{n,t+1} (I_n - (\hat{g}_{2,n,t+1}^\prime \hat{g}_{2,n,t+1} - \hat{g}_{2,n,t+1}^\prime \hat{g}_{2,n,t+1})\hat{g}_{2,n,t+1})
\]

(5.10)

(5.11)

as well as \(a_{n,T}(\beta_0) = 0\) and for \(t = T - 1, \ldots, 1\):

\[
a_{n,t}(\beta_0) = a_{n,t+1}(\beta_0)
\]

\[
+ B_{n,t+1}(\hat{g}_{2,n,t}^\prime \hat{g}_{2,n,t}) \dddot{g}_{2,n,t}(\dot{y}_{n,t} - \dot{x}_{n,t}\beta_0 - \ddot{g}_{2,n,t}a_{n,t+1}(\beta_0))
\]

(5.12)

Then, by recursion, the information for estimating \(\beta_0\) contained in (2.3) and (2.4) is equivalent to the information contained in:

\[
E(M_{\hat{g}_{2,n,t}}(\dot{y}_{n,t} - \dot{x}_{n,t}\beta_0 - \ddot{g}_{2,n,t}a_{n,t+1}(\beta_0)) | Z_t) = 0 \forall t = 1, \ldots, T - 1
\]

(5.13)

Define \(y^n_t = [y_{n,t}]_{t' = T, \ldots, t}; \ x^n_t = [x_{n,t}]_{t' = T, \ldots, t}; \ g^n_t = [g_{n,t}]_{t' = T, \ldots, t}.\)
Appendix D shows that (5.13) is equivalent to:

\[
E(y_{n,t} - x_{n,t} \beta_0 - g_{n,t}(g_n' g_n^{-1}) - g_n'(y_{n,t} - x_{n,t} \beta_0)|Z_t) = 0 \forall t = 1, ..., T - 1
\]

(5.14)

The moment conditions in (5.14) present the important advantage over the moment conditions in (5.13) of being serially uncorrelated when transitory shocks are homoscedastic and serially and cross-sectionally uncorrelated. As a result locally efficient estimators are easier to define starting from (5.14) instead of applying a sequence of orthogonalizations to (5.13) as in Chamberlain [1992] which in the case of two-way fixed effects considered here would quickly become intractable.

As in Section 3, it is possible for \( ||g_{n,t}(g_n' g_n^{-1}) - g_n'^t ||_\infty \) to grow unboundedly as \( n \) grows when \( N_{n,d} \leq C < \infty \ \forall \ d \in D_n \). Therefore, as in Section 3, we will apply the transformation used in (5.14) to many subsets of the data chosen such that \( d_{it} \) only takes finitely many values in each subset. In the next section we define an estimator for \( \beta_0 \) based on these moment conditions.

5.2 Estimator and Asymptotic Properties

Define \( S_n \) to be a collection of subset of \( \mathcal{N} \) as in Section 3. For \( \sigma \in S_n \), define \( x_{n,t}^{t(\sigma)} = [x_{it}]_{i \in \sigma, t = T, ..., t}, y_{n,t}^{t(\sigma)} = E(x_{n,t}^{t(\sigma)} | Z_t), \ g_{n,t}^{(\sigma)} = [\{1[i = i^t] \}_{i \in \sigma}], [d_{it}]_{i \in \sigma, t = T, ..., t}, \ g_{n,t}^{t(\sigma)} = \Sigma_{n,t}^{(\sigma)} = (I_{\# \sigma} - g_{n,t}^{t(\sigma)} g_{n,t}^{t(\sigma)'} g_{n,t}^{t(\sigma)})^{-1} \ Sigma_{n,t}^{(\sigma)} (5.15) \)

The estimator we study in this section is:

\[
\hat{\beta}_n = \left( \sum_{\sigma \in S_n} \sum_{t} \tilde{z}_{n,t}^{t(\sigma)} M_{n,t}^{(\sigma)} \Sigma_{n,t}^{(\sigma)} M_{n,t}^{(\sigma)'} \right)^{-1}
\]

\[
\times \sum_{\sigma \in S_n} \sum_{t} \tilde{z}_{n,t}^{t(\sigma)} M_{n,t}^{(\sigma)} \Sigma_{n,t}^{(\sigma)} M_{n,t}^{(\sigma)'} y_{n,t}^{t(\sigma)}
\]

(5.17)

Define \( \tilde{z}_{n,t}^{t(\sigma)} \) to be the last \( \# \sigma \times dim(x_{it}) \) block of \( \tilde{z}_{n,t}^{t(\sigma)} \), \( \tilde{z}_{n,t+1}^{t(\sigma)} \) to be the first \( \# \sigma (T - t) \times dim(x_{it}) \) block of \( \tilde{z}_{n,t+1}^{t(\sigma)} \).
\[ \hat{\beta}_n = \left( \sum_t \hat{Z}_{t,n,t} x_{n,t} \right)^{-1} \sum_t \hat{Z}_{t,n,t} y_{n,t} \]
Assumption 5, in addition to imposing mean independence of transitory shocks across observations with different values for $d_{it}$ in a given time period, imposes mean independence with respect to past shocks. In the previous section we could invoke a central limit theorem conditional on $\{Z_t\}_{t \in T}$ as part of the proof of asymptotic normality of the estimator under Assumption 3 of cluster independence of the instrumental variables. Here this is not possible since the instrument set varies over time. Instead we rely on a central limit theorem for martingale difference sequences, which then necessitates mean independence of the transitory shocks over time.\footnote{Here we do not pursue estimation methods that use the additional moment conditions arising from the assumption of serial uncorrelation for efficiency, as in Ahn and Schmidt (1995), Arellano and Bover (1995), or Blundell and Bond (1998) for models with one-way fixed effects. The resulting estimators would be inconsistent if the assumption of serial uncorrelation failed, while the estimator we propose is easily shown to be consistent under serial correlation. In addition, one can show asymptotic normality of the estimator proposed here in the presence of serial correlation by assuming, for instance, that $u_{it}$ is composed of two independent terms, one satisfying Assumption 1, the other satisfying Assumption 5. This is not pursued here for simplicity.}

The next Assumption imposes restrictions on moments of the data.

**Assumption 6.** There is a constant $c > 0$ and random variables $B_{it} \in \sigma(R_{it})$ with $E(B_{it}) \leq C < \infty$ $\forall \{i, t\} \in N \times T$ such that:

- $\lambda_{\min}(\frac{1}{n}E(\sum_{t} z'_{n,t} x_{n,t})) \geq c \forall n \geq N$.
- $V_{t} = \lim_{n \to \infty} \frac{1}{n} \text{Var}(\hat{z}'_{n,t} u_{n,t})$ exists and is positive definite $\forall t \in T$.
- $E(||u_{it}||^{4+\delta}|R_{it}) \leq B_{it}$, $E(||x_{it}||^{4+\delta}|R_{it}) \leq B_{it}$ a.e. $\forall \{i, t\} \in N \times T$.

This assumption is very similar to Assumption 2 in the previous section, except that it imposes that the limit of $\frac{1}{n} \text{Var}(\hat{z}'_{n,t} u_{n,t})$ exists, which will be needed for showing that $\sqrt{n}(\hat{\beta}_{n} - \beta_{0})$ is asymptotically normal in this section.

We modify slightly Assumptions 3 and 4 stated in Section 4 for this section.

**Assumption 7.** $\forall n$, there exists a partition $P_{n}$ of $N$ such that $\{\{d : d_{it}, i \in p, t \in T\}, p \in P_{n}\}$ forms a partition of $D_{n}$, and $\forall p, p' \in P_{n}$, $p \neq p'$, $F(\{x_{it}, z_{it}, u_{it}\}_{i \in p, t \in T}, \{x_{it}, z_{it}, u_{it}\}_{i \in p', t \in T}) = F(\{x_{it}, z_{it}, u_{it}\}_{i \in p, t \in T})$ and $\frac{\#p}{\#p'} \leq C < \infty$. $\#P_{n} \to \infty$ as $n \to \infty$.

**Assumption 8.** Assumption 4 holds but for Assumption 4-c, which is replaced by $\forall i, j \in N$ s.t. $d_{it} = d_{jt}$:
\[ E(u_{it}u_{jt}|R_{it}) = E(u_{it}u_{jt}|\{z_{i't}\}_{d_{it}=d_{i't}}) \quad (5.21) \]

Assumption 8 restricts the heteroscedasticity in the variance-covariance matrix of \( u_{it} \) so that it depends on \( Z_t \) only, and not on \( \{u_{is} : s < t \text{ or } s = t, d_{jt} \neq d_{it}\} \). Additionally the dependence on \( Z_t \) is restricted to past observations or observations that share the same value of \( d_{it} \), as in Assumption 4-c.\(^{14}\)

Under the assumptions above, we can show that the estimator proposed here is consistent and asymptotically normal.

**Proposition 4.** Under (2.3) and (2.4) with \( Z_t \subseteq Z_s \) for \( s > t \), Assumptions 5, 6, and Assumption 7 or Assumption 8, as \( n \to \infty \) while \( N_{n,d} \leq C \forall d \in D_n, \#\{d : d_{it} = d, i \in \sigma, t \in T\} \leq C \forall \sigma \in S_n, \) and \( \#\{\sigma \in S_n : i \in \sigma\} \leq C \forall \sigma \in S_n, C < \infty, \forall n:\)

\[
\hat{\beta}_n \overset{p}{\to} \beta_0 \quad (5.22)
\]

and:

\[
V^{-\frac{1}{2}} A_n \sqrt{n}(\hat{\beta}_n - \beta_0) \overset{d}{\to} N(0, I) \quad (5.23)
\]

where \( A_n = \frac{1}{n} E(\sum_t z_{n,t}'x_{n,t}) \) and \( V = \sum_{t=1}^{T} V_t. \)

We can define consistent standard errors for the estimator proposed here as in Section 4.

Let \( S_n \) and \( \hat{u}_n \) be defined as in Section 4.

Redefine \( \tilde{z}_n = [\tilde{z}_{n,t}]_{t=T-1,...,1}. \)

For any \( \sigma \in S_n \), define:

\[
S_{n,t}^{(\sigma)} = [1[i = j, s = t']]_{i \in \mathcal{N}, s \in T}_{j \in \sigma, t' = T, ..., t} \quad (5.24)
\]

and redefine:

\[
S_n^{(\sigma)} = [S_{n,t}^{(\sigma)}]_{t=T-1,...,1} \quad (5.25)
\]

\(^{14}\)Note that \( z_{it-1} \subseteq z_{it} \), so that \( z_{it} \) includes previous realizations of the instrumental variables.
Define:

\[ M_{n,t}^{2(\sigma)} = g_{n,t+1}(\sigma)^\prime n - (g_{n,t}(\sigma)^\prime n)g_{n,t+1}(\sigma)^\prime n - g_{n,t}(\sigma)^\prime n - I_{\#\sigma} - g_{n,t}(\sigma)^\prime n - g_{n,t}(\sigma)^\prime n - g_{n,t}(\sigma)^\prime n \]

and:

\[ M_n^{2(\sigma)} = \text{diag}(\{ M_{n,t}^{2(\sigma)} \}) \]

Define:

\[ M_n^2 = [S_n^{(\sigma)}]^\prime \sigma \in S_n \text{diag}(\{ M_n^{2(\sigma)} \}) [S_n^{(\sigma)}] \sigma \in S_n \]

so that we have \( \hat{\beta}_n = (\tilde{z}_n M_n^2 x_n)^{-1} \tilde{z}_n M_n^2 y_n \).

**Proposition 5.** Under (2.3) and (2.4) with \( Z_t \subseteq Z_s \) for \( s > t \), Assumptions 5, 6, and Assumption 7 or Assumption 8, as \( n \to \infty \) while \( N_{n,d} \leq C \forall d \in D_n, \#\{d : d_{it} = d, i \in \sigma, t \in T\} \leq C \forall \sigma \in S_n, \) and \( \#\{\sigma \in S_n : i \in \sigma\} \leq C \forall i \in N, C < \infty, \forall n \):

\[ \frac{1}{n} \sum_{t} \tilde{z}_{n,t}^\prime x_{n,t} - A_n \xrightarrow{p} 0 \]  
\[ \frac{1}{n} \tilde{z}_n (M_n^2 \tilde{u}_n M_n^2 \odot S_n) \tilde{z}_n - V \xrightarrow{p} 0 \]

Under homoscedasticity and serial and cross-sectional uncorrelation, we can also show that the proposed estimator is asymptotically efficient in the class of instrumental variable estimators when, as in Section 4, \( S_n \) is a partition of \( N \).

**Proposition 6.** If \( S_n \) is a partition of \( N \), \( \{ \{d_{it}\}_{i \in \sigma, t \in T} \}_{\sigma \in S_n} \) is a partition of \( D_n \) and if

\[ \text{Cov}(u_{it}, u_{js} | Z_{\max\{t,s\}}) = \begin{cases} \sigma_u^2 & \text{if } i = j, t = s \\ 0 & \text{otherwise} \end{cases} \]

then:

\[ B_n^{-1}W_n B_n^{-1} - A_n^{-1}V_n A_n^{-1} \]

is positive semi-definite \( \forall n \), where \( B_n = \frac{1}{n} E(\sum_t \tilde{z}_{n,t}^\prime (x_{n,t} - g_{n,t}(\sigma_n^\prime n)g_{n,t}^\prime n)) \) and \( W_n = \frac{1}{n} E(\sum_t \tilde{z}_{n,t}^\prime (x_{n,t} - g_{n,t}(\sigma_n^\prime n)g_{n,t}^\prime n)) (5.32) \)
\[ \frac{1}{n} \text{Var}(\sum_t \hat{z}_{n,t}^t(u_{n,t} - g_{n,t}(g_{n,t}^t - g_{n,t}^t u_{n,t}))) \] and \( \hat{z}_{n,t} \) is any \( n \times \text{dim}(\beta_0) \) function of \( \{z_{it}\}_{i \in \mathbb{N}} \).

The proofs for Propositions 4-6 are found in Appendix D.

### 6 Monte Carlo Simulation Study

In this section I present the results from a brief Monte Carlo simulation study that looks at estimating a dynamic model with one strictly exogenous covariate and two-way fixed effects. I study the small sample properties for the estimator proposed in section 5 and alternative estimators for \( \rho_0 \) and \( \beta_0 \) in the model given by, \( \forall t = 2, \ldots, T \):

\[
y_{it} = \rho_0 y_{i(t-1)} + \beta_0 x_{it} + c_i + e_{dit} + u_{it} \tag{6.1}
\]

\[
E(u_{it}|Y^{t-1}, X) = 0 \tag{6.2}
\]

where \( Y^{t-1} = [y_{is}]_{i=1,n,s=1,t-1} \).

In order to implement the estimator defined in Section 5.2 by (5.20), I replace \( E(y_{it-1}|Y^{s-1}, X) \) \( \forall t = 2, \ldots, T, s \leq t \) in the formula given in (5.20) by the estimated predicted values from a regression of \( y_{it-1} \) on \( s_1 = y_{is-1}, s_2 = \frac{1}{s-t+1} \sum_{\tau=2}^s y_{i\tau-1}, s_3 = \frac{1}{\#\{j:d_{js}=d_{is}\}} \sum_{j:d_{js}=d_{is}} y_{js-1}, s_4 = \frac{1}{\#\{j:d_{js}=d_{is}\}} \sum_{j:d_{js}=d_{is}} s_{jjs}, x_i = \{x_{i\tau}\}_{\tau=1,T}^{15} \). \( \mathcal{S}_n \) is defined to be \( \{j \in \mathbb{N}: d_{jt} = d_{is}, t, s = 2, \ldots, T\} \) \( \forall i \in \mathbb{N} \). I call this estimator the dynamic two-way fixed effects estimator (DTW).

The first alternative estimator considered is the estimator proposed in section 4, which treats \( y_{it-1} \) as a strictly exogenous covariate. I call this estimator the static two-way fixed effects estimator (STW).

The second alternative estimator is an estimator as in Section 5.2 but that treats \( [1[d_{it} = d]]_{d \in \mathcal{D}_n} \) as a finite dimensional vector of strictly exogenous covariates, so that only \( c_i \) is removed by sequential transformations of the data as in Section 5.2. I call this estimator the naive two-way fixed effects estimator (NTW).

---

\[ ^{15} \text{With the dynamic model used in this section, it is possible to develop an auxiliary model for } E(y_{it-1}|Y^{s-1}, X) \text{ that could be both more parsimonious and a better predictor in the presence of cross-sectional dependence than the sequence of linear projections used here, see } \text{Arellano (2016) and Verdier (2016). This is not considered here for simplicity.} \]
The third alternative estimator is as in section 5.2 but ignores \( e_{d_{it}} \) and estimates this dynamic model with cross-sectional unobserved heterogeneity only. I call this estimator the one-way fixed effects estimator 1 (OW1).

The fourth alternative estimator is as in section 5.2 but ignores \( c_i \) and estimates the model with unobserved heterogeneity indexed by values of \( d_{it} \) only. I call this estimator the one-way fixed effects estimator 2 (OW2).

NTW, OW1, and OW2 all rely on using the same predictors as DTW, \( s_{is}^1, s_{is}^2, s_{is}^3, s_{is}^4, x_i \), in order to compute exactly identifying instruments for their respective transformed covariates.

Finally, the fifth alternative estimator ignores all forms of unobserved heterogeneity and estimates \( \rho_0 \) and \( \beta_0 \) by a regression of \( y_{it} \) on \( y_{it-1} \) and \( x_{it} \) (OLS).

The data generating process used for this Monte Carlo analysis uses the values: \( \rho_0 = 0.5 \), \( \beta_0 = 1 \), \( n = 250 \), \( N_n = 50 \), \( T = 5 \), and:

\[
\begin{align*}
    e_d & \sim \chi^2_{3,c} \\
u_{i0} & \sim \chi^2_{3,c} \\
w_{it} & \sim \chi^2_{3,c} \\
v_{it} & \sim \chi^2_{3,c} \\
u_{it} & \sim \chi^2_{3,c}
\end{align*}
\]

where \( \chi^2_{3,c} \) is the distribution obtained by standardizing the \( \chi^2 \) distribution with three degrees of freedom. Random variables are drawn independently wherever dependence is not shown explicitly.

(6.3) describes the process that assigns values of \( d_{it} \) to a particular cross-sectional observation \( i \) at time \( t \). It states that groups of five observations will be assigned to the value of \( e_d \) that corresponds to their rank in terms of the sum of the unobserved heterogeneity term \( c_i \), initial unobserved shocks \( u_{i0} \), contemporaneous shock to the strictly exogenous covariate \( w_{it} \), and an assignment shock \( v_{it} \). This is a simple assignment rule designed to illustrate non-random assignment of values of \( d_{it} \), for instance assignment of students to teachers based on [6.6].
on student unobserved factors and teacher unobserved quality.

In Table 1 I report summary statistics from the simulated distributions of the estimators of $\rho_0$ and $\beta_0$: the expected deviation of the estimators from the true values of the parameters, the expected absolute deviation, the median deviation, and several quantiles of the deviation of the estimators from the median of their distributions. 10,000 replications were used to obtain the simulated distributions of the estimators.

Looking at bias, median deviation, or mean absolute deviation of the estimators, we see that it is not only accounting for both sources of heterogeneity that is important for estimating both $\rho_0$ and $\beta_0$ accurately (DTW outperforms OW1, OW2, and OLS), but also accounting for the fact that one of the covariates is sequentially exogenous (DTW outperforms STW), and using an estimator that treats neither $c_i$ nor $e_d$ as consistently estimated (DTW outperforms NTW). As with models with one-way fixed effects only, the large reduction in bias exhibited by the DTW estimator can come at the cost of an increase in sampling variation compared to estimators that rely on the original data (OLS) or partially transformed data (OW2), or compared to estimators that treat covariates as strictly exogenous instead of using an instrumental variable approach (STW).

Appendix E contains additional results for various parameterizations of this data generating process.

7 The Effect of Class-Size Reductions on Student Achievement

In this section I estimate the effects of class size on student achievement among elementary school students in North Carolina, using a dynamic model of learning that accounts for non-random sorting based on student unobserved ability and teacher and school unobserved quality.

Class size reduction has been a policy of interest in education for several decades now, and many studies have been dedicated to evaluating its effectiveness. A review of the literature can be found, among others, in Hanushek (1997), Hanushek (2006), Whitehurst and Chingos.
Three main methodological approaches have been used to evaluate the effectiveness of class size reduction. The first relies on experimental evidence using data where students were randomly assigned to classrooms of different sizes (e.g. Krueger (1999), De Giorgi et al. (2012), Duflo et al. (2015)). The second relies on quasi-experimental evidence using legal or common practice upper bounds on class size and the size of incoming cohorts as natural experiments (e.g. Angrist and Lavy (1999), Hoxby (2000)). The third relies on the use of longitudinal data (e.g. Hoxby (2000), Rivkin et al. (2005), Bandiera et al. (2010)).

The data used here is student level longitudinal data made available by the North Carolina Education Research Center. The first part of the data contains scores on standardized tests in mathematics and reading for individual students in each grade. The second part of the data contains course enrollment information and instructor identification, so that students can be linked to their teachers.

I restrict the estimation sample to students in elementary schools (grades three to five) because the structure of their course enrollment is simpler than students in middle or high school and I exclude charter schools because instructor identification seems less reliable for observations in charter school. I also use a sample made of the years 2008 to 2012. 2012 is the last year available at this time. The year 2008 is chosen as start year because the version of the mathematics and reading tests does not change from 2008 to 2012, whereas different versions of the tests were used prior to 2008, and equating scores across years is non-trivial and could invalidate the results presented here. I restrict the estimation sample to students who follow a normal succession of grades (do not repeat or skip grades) and who are observed in the data for three consecutive grades. In order to avoid small sample issues for small class sizes, I restrict the estimation sample to students enrolled in classes of between 10 and 30 students.

Table 2 shows several statistics from the sample used for estimation, such as unconditional average and standard deviation of test scores, and average test scores for various values of

---

16The code that produced the dataset used here is available on the author’s website. Paul Hanselman provided very graciously code to process the data on course enrollment, which was used in Hanselman (2014).
class size. Test scores are left in their original units here. These units can be contextualized by considering percentiles of the distribution of test scores. An increase in test score by one would switch a 50th percentile student to the 54th percentile in mathematics and a 48th percentile student to the 52nd percentile in reading. On average an increase in test score by one is associated with moving up in the grade distribution by around three percentiles. One pattern to highlight in Table 2 is that students tend to have higher test scores in larger classes, a pattern which is still present after controlling for grade-year systematic differences as shown in the last column of Tables 5 and 6. Figure 1 shows average test scores by class size for all values of class size in the estimation sample. Tables 3 and 4 present observed frequencies of transition for class size from grade 4 to grade 5 for students and from years 2009 to 2010, 2010 to 2011, 2011 to 2012 for teachers and shows that there is significant movement of students and teachers across different class sizes, which is the basis for the rank conditions stated in Sections 4 and 5 to hold. Figure 2 shows a visualization of the largest subcomponent of $G_t$ in mathematics and reading, which contains more than 99% of all teachers in both subjects.

The outcomes of interest are student achievement in mathematics and reading as measured by test scores. These tests are scored according to item response theory scoring\[^{17}\] which aims at producing test scores that can be compared not only cross-sectionally across students in the same grade, but also over time for students in different grades. In Appendix F I show results that correspond to a model with classical measurement error in test scores. The joint estimation of a model of non-classical measurement error\[^{18}\] with the dynamic model of learning used here is left for future work.

The policy of interest is reductions in the number of students in classes. Anecdotal evidence obtained from conversations with teachers as well as a prior towards decreasing marginal returns to teacher time per pupil lead to considering non-linear effects of class size on learning. To accommodate this potential non-linearity, I use a polynomial of order four to model the response function of student achievement to class size.\[^{19}\]

\[^{17}\]See e.g. Thissen and Wainer (2001).
\[^{18}\]See e.g. Nielsen (2015).
\[^{19}\]Note that using the natural logarithm of class size as a covariate would lead to the opposite non-linearity compared to what is expected since it would imply that the effect of a reduction in class size by a fixed
The preferred estimate of the effect of class size on student achievement here will rely on a dynamic model of learning, so that previous educational inputs have an effect on current learning, which is assumed to decay at a geometric rate (see e.g. [Todd and Wolpin (2003), Andrabi et al. (2011)]). The model will also account for student unobserved ability, teacher unobserved quality, and school unobserved quality, in the form of additively separable unobserved heterogeneity as described in the rest of the paper. Finally grade-year specific intercepts are included in the model and treated as coefficients on finitely dimensional covariates.

The resulting model can be written:

\[ y_{it} = \alpha_{\text{grade}_{it}, t} + \rho_0 y_{it-1} + x_{it}/\beta_0 + c_i + e_{dit} + f_{\text{school}_{it}} + u_{it} \]  
(7.1)

\[ E(u_{it}|Y^{t-1}, X) = 0 \]  
(7.2)

where \( y_{it} \) is student \( i \)'s test score in year \( t \), either in mathematics or reading, \( x_{it} = [(cs_{it})^k]_{k=1,...,4} \) where \( cs_{it} \) is the number of students in student \( i \)'s classroom in year \( t \), \( d_{it} \) is student \( i \)'s teacher in year \( t \), \( \text{school}_{it} \) is student \( i \)'s school in year \( t \), and \( \text{grade}_{it} \) is student \( i \)'s grade in year \( t \).

The parameters of this model, and the corresponding standard errors, are estimated as described in Section 5. The details of the estimation methods used in this section are discussed in Appendix F.

In order to evaluate the difference between the results obtained with the approach outlined in this paper compared to existing methods, I also present results obtained with several alternative estimators.

For dynamic models, I show results obtained with an estimator for dynamic models with one-way unobserved heterogeneity indexed by student identity only, as discussed for instance in [Andrabi et al. (2011), Rothstein (2010) and Verdier (2016)]. I also show results obtained with an estimator for models without student level unobserved heterogeneity and which treats number of students is lower when starting from a large class size than starting from a small class size, whereas decreasing marginal returns to teacher time per pupil would imply the opposite.

Note that here I use a model with three-way unobserved heterogeneity instead of two-way only. The results in Sections 4 and 5 are easily shown to extend to the case of three-way unobserved heterogeneity when using an asymptotic framework where few teachers are assigned to each school, which seems appropriate for this empirical application since on average only around ten teachers teach grades 4 and 5 in each school.
teacher and school level unobserved heterogeneity as finite dimensional parameters that can be estimated consistently. This type of estimator is used for instance in Kane and Staiger (2008), Chetty et al. (2014), Guarino et al. (2014). A heuristic argument frequently found in the empirical literature for using estimators that correspond to dynamic models without student level unobserved heterogeneity is that past test scores serve as proxies for student unobserved ability. Finally I show results obtained with an estimator for models without unobserved heterogeneity, which is simply an ordinary least squares regression of test score on past test score, grade-year indicators, and the polynomial in class size.

I also consider estimators for a model in gains ($\rho_0$ is assumed to equal one) where past test score is not included as an explanatory variable but the dependent variable is replaced by the first difference in test score. This type of model is also frequently used in empirical work, e.g. Rivkin et al. (2005), Kane and Staiger (2008). For this model I consider the same four configurations of sources of unobserved heterogeneity: Student, teacher and school level unobserved heterogeneity which leads to using an estimator as in Section 4 of this paper; Student level only, which leads to using a simple one-way fixed effects estimator; Teacher and school level unobserved heterogeneity treated as finite dimensional covariates that can be estimated consistently; No unobserved heterogeneity which leads to an ordinary least squares regression of the first difference in test score on grade-year fixed indicators and the polynomial in class size.

Finally I also consider the same estimators as for the model in gains but for a static model of learning ($\rho_0$ is assumed to equal zero), this type of model is used for instance in Bandiera et al. (2010).

The results are presented in Tables 5 and 6, and additional visualizations of the results are shown in Figures 3-5.

Firstly in mathematics, on a range of class size from 15 students to 30 students, reductions in class size are estimated to have a positive effect on student achievement, and the evolution of the magnitude of this effect over the range of class size corresponds to the hypothesis

---

21 When past test score is not included as an explanatory variable, the assumptions of the model imply that all remaining covariates are strictly exogenous.
of decreasing marginal returns to teacher time per pupil. The magnitude of this effect is relatively modest, e.g. reducing class size from 30 to 15 students is estimated to lead to an average increase in test score of around 0.8, i.e. around 10% of a standard deviation, or a move up in the distribution of test scores of around 3 percentiles on average.

In reading reductions in class size are estimated not to have a statistically significant effect on student achievement. In mathematics, on a range of class size from 10 to 15 students, reductions in class size are estimated to lead to a decrease in student achievement. This reversal in the sign of the effect of class size reduction could be due to factors that are not controlled for here but are related to class size. In particular ongoing work is interested in how these results could change when accounting for peer effects.

Secondly it appears from these results that there is positive selection not only of students into larger class sizes, but also of teachers, since estimators that do not account for student and teacher unobserved heterogeneity simultaneously lead to smaller estimates of the effect of class size reduction on student achievement than the estimator proposed in this study (Figure 4).

Finally, in mathematics, estimators that correspond to a model in gains lead to estimated effects of class size reductions that are significantly larger than results obtained with a dynamic model while the difference with the results obtained with estimators that correspond to a static model is small (Figure 5). This corresponds to the estimate of $\rho_0$ being small in magnitude when estimating a dynamic model. In reading, $\rho_0$ is estimated to be higher, but since class size reduction is not estimated to have a significant effect on achievement, there is no large difference between using a model that is dynamic, in gains, or static. Note that the results presented in Appendix F that correspond to using a model with classical measurement error lead to larger estimates for $\rho_0$, so that in mathematics there is a larger difference in the results obtained with a dynamic model and with a static model.
8 Conclusion

This paper provides an asymptotic justification for estimation and inference with linear models of panel data that include multi-way unobserved heterogeneity even when few observations are assigned to each values of the variables that index unobserved heterogeneity and when cross-sectional dependence is present. It also defines a new estimator for dynamic models (models with sequentially exogenous instruments), so that practitioners do not have to make a choice between including one-way unobserved heterogeneity only, suppressing the dynamics of the model, or treating all but one term of unobserved heterogeneity as finite dimensional parameters that can be estimated consistently.

When applied to test score data from North Carolina, the methods developed in this paper seem to be useful for evaluating the effect of a policy such as class size reduction on student achievement, yielding statistically significant positive effects of reducing class size on student achievement in Mathematics which can be compared with the results obtained with estimators for models with one-way unobserved heterogeneity only or that suppress dynamics.

References


Table 1: Results from Monte Carlo simulations for different estimators of $\rho_0 = .5$ and $\beta_0 = 1.$

<table>
<thead>
<tr>
<th></th>
<th>DTW</th>
<th>STW</th>
<th>NTW</th>
<th>OW1</th>
<th>OW2</th>
<th>OLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E(\hat{\rho} - \rho_0)$</td>
<td>-0.038</td>
<td>-0.277</td>
<td>-0.128</td>
<td>-0.158</td>
<td>0.276</td>
<td>0.257</td>
</tr>
<tr>
<td>$E(</td>
<td>\hat{\rho} - \rho_0</td>
<td>)$</td>
<td>0.062</td>
<td>0.277</td>
<td>0.128</td>
<td>0.160</td>
</tr>
<tr>
<td>$Med(\hat{\rho} - \rho_0)$</td>
<td>-0.038</td>
<td>-0.277</td>
<td>-0.126</td>
<td>-0.159</td>
<td>0.277</td>
<td>0.258</td>
</tr>
<tr>
<td>$Q_{0.05}(\hat{\rho} - Med(\hat{\rho}))$</td>
<td>-0.110</td>
<td>-0.078</td>
<td>-0.102</td>
<td>-0.130</td>
<td>-0.077</td>
<td>-0.055</td>
</tr>
<tr>
<td>$Q_{0.25}(\hat{\rho} - Med(\hat{\rho}))$</td>
<td>-0.045</td>
<td>-0.031</td>
<td>-0.041</td>
<td>-0.054</td>
<td>-0.030</td>
<td>-0.022</td>
</tr>
<tr>
<td>$Q_{0.75}(\hat{\rho} - Med(\hat{\rho}))$</td>
<td>0.045</td>
<td>0.032</td>
<td>0.039</td>
<td>0.055</td>
<td>0.029</td>
<td>0.021</td>
</tr>
<tr>
<td>$Q_{0.95}(\hat{\rho} - Med(\hat{\rho}))$</td>
<td>0.111</td>
<td>0.078</td>
<td>0.092</td>
<td>0.136</td>
<td>0.069</td>
<td>0.051</td>
</tr>
<tr>
<td>$E(\hat{\beta} - \beta_0)$</td>
<td>-0.029</td>
<td>-0.231</td>
<td>-0.104</td>
<td>-0.774</td>
<td>0.328</td>
<td>-0.695</td>
</tr>
<tr>
<td>$E(</td>
<td>\hat{\beta} - \beta_0</td>
<td>)$</td>
<td>0.133</td>
<td>0.239</td>
<td>0.153</td>
<td>0.774</td>
</tr>
<tr>
<td>$Med(\hat{\beta} - \beta_0)$</td>
<td>-0.028</td>
<td>-0.230</td>
<td>-0.102</td>
<td>-0.769</td>
<td>0.327</td>
<td>-0.690</td>
</tr>
<tr>
<td>$Q_{0.05}(\hat{\beta} - Med(\hat{\beta}))$</td>
<td>-0.271</td>
<td>-0.259</td>
<td>-0.266</td>
<td>-0.352</td>
<td>-0.244</td>
<td>-0.287</td>
</tr>
<tr>
<td>$Q_{0.25}(\hat{\beta} - Med(\hat{\beta}))$</td>
<td>-0.109</td>
<td>-0.103</td>
<td>-0.110</td>
<td>-0.140</td>
<td>-0.096</td>
<td>-0.111</td>
</tr>
<tr>
<td>$Q_{0.75}(\hat{\beta} - Med(\hat{\beta}))$</td>
<td>0.111</td>
<td>0.105</td>
<td>0.106</td>
<td>0.135</td>
<td>0.100</td>
<td>0.106</td>
</tr>
<tr>
<td>$Q_{0.95}(\hat{\beta} - Med(\hat{\beta}))$</td>
<td>0.266</td>
<td>0.253</td>
<td>0.260</td>
<td>0.323</td>
<td>0.244</td>
<td>0.257</td>
</tr>
</tbody>
</table>

DTW is the estimator proposed in section 2.2. STW treats covariates as strictly exogenous, NTW treats one of the fixed effects as finite-dimensional, OW1 and OW2 each ignore one of the fixed effects, OLS ignores both fixed effects.
Table 2: Effect of class size on student achievement, description of the estimation sample.

<table>
<thead>
<tr>
<th></th>
<th>Mathematics</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
<td>171,377</td>
<td></td>
</tr>
<tr>
<td>Teachers</td>
<td>11,696</td>
<td>12,057</td>
</tr>
<tr>
<td>Years</td>
<td>2009 to 2012</td>
<td></td>
</tr>
<tr>
<td>Grades</td>
<td>4 and 5</td>
<td></td>
</tr>
<tr>
<td>Test Scores</td>
<td></td>
<td></td>
</tr>
<tr>
<td>average</td>
<td>355.12</td>
<td>349.25</td>
</tr>
<tr>
<td>standard dev.</td>
<td>9.37</td>
<td>9.50</td>
</tr>
<tr>
<td>Class size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>average</td>
<td>22.03</td>
<td>21.61</td>
</tr>
<tr>
<td>standard dev.</td>
<td>4.07</td>
<td>4.44</td>
</tr>
<tr>
<td>Average test score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>by class size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-14 students</td>
<td>353.61</td>
<td>347.91</td>
</tr>
<tr>
<td>15-19 students</td>
<td>353.22</td>
<td>347.38</td>
</tr>
<tr>
<td>20-24 students</td>
<td>354.87</td>
<td>349.06</td>
</tr>
<tr>
<td>24-30 students</td>
<td>357.00</td>
<td>351.23</td>
</tr>
</tbody>
</table>
Table 3: Observed frequencies of transition between class sizes, Mathematics.

<table>
<thead>
<tr>
<th></th>
<th>10-14</th>
<th>15-19</th>
<th>20-24</th>
<th>25-30</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students, grade 4 (left) to grade 5 (top).</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-14</td>
<td>966</td>
<td>2,235</td>
<td>5,010</td>
<td>3,574</td>
<td>11,785</td>
</tr>
<tr>
<td>15-19</td>
<td>1,992</td>
<td>10,043</td>
<td>17,313</td>
<td>7,245</td>
<td>36,593</td>
</tr>
<tr>
<td>20-24</td>
<td>2,757</td>
<td>10,799</td>
<td>43,558</td>
<td>24,379</td>
<td>81,493</td>
</tr>
<tr>
<td>25-30</td>
<td>1,169</td>
<td>2,872</td>
<td>13,988</td>
<td>23,477</td>
<td>41,506</td>
</tr>
<tr>
<td>Total</td>
<td>6,884</td>
<td>25,949</td>
<td>79,869</td>
<td>58,675</td>
<td>171,377</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>10-14</th>
<th>15-19</th>
<th>20-24</th>
<th>25-30</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-14</td>
<td>111</td>
<td>229</td>
<td>304</td>
<td>159</td>
<td>803</td>
</tr>
<tr>
<td>15-19</td>
<td>146</td>
<td>820</td>
<td>1,126</td>
<td>361</td>
<td>2,453</td>
</tr>
<tr>
<td>20-24</td>
<td>200</td>
<td>812</td>
<td>2,516</td>
<td>1,226</td>
<td>4,754</td>
</tr>
<tr>
<td>25-30</td>
<td>73</td>
<td>202</td>
<td>979</td>
<td>1,183</td>
<td>2,437</td>
</tr>
<tr>
<td>Total</td>
<td>530</td>
<td>2,063</td>
<td>4,925</td>
<td>2,929</td>
<td>10,447</td>
</tr>
</tbody>
</table>
Table 4: Observed frequencies of transition between class sizes, Reading.

<table>
<thead>
<tr>
<th>Students, grade 4 (left) to grade 5 (top).</th>
<th>10-14</th>
<th>15-19</th>
<th>20-24</th>
<th>25-30</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-14</td>
<td>3,183</td>
<td>4,027</td>
<td>7,550</td>
<td>4,423</td>
<td>19,183</td>
</tr>
<tr>
<td>15-19</td>
<td>2,662</td>
<td>10,007</td>
<td>17,107</td>
<td>7,223</td>
<td>36,999</td>
</tr>
<tr>
<td>20-24</td>
<td>3,164</td>
<td>10,453</td>
<td>39,895</td>
<td>22,179</td>
<td>75,691</td>
</tr>
<tr>
<td>25-30</td>
<td>1,229</td>
<td>2,791</td>
<td>12,896</td>
<td>22,588</td>
<td>39,504</td>
</tr>
<tr>
<td>Total</td>
<td>10,238</td>
<td>27,278</td>
<td>77,448</td>
<td>56,413</td>
<td>171,377</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10-14</td>
<td>365</td>
<td>422</td>
<td>660</td>
<td>259</td>
<td>1,706</td>
</tr>
<tr>
<td>15-19</td>
<td>196</td>
<td>744</td>
<td>1,055</td>
<td>349</td>
<td>2,344</td>
</tr>
<tr>
<td>20-24</td>
<td>214</td>
<td>704</td>
<td>2,075</td>
<td>1,087</td>
<td>4,080</td>
</tr>
<tr>
<td>25-30</td>
<td>68</td>
<td>169</td>
<td>825</td>
<td>1,069</td>
<td>2,131</td>
</tr>
<tr>
<td>Total</td>
<td>843</td>
<td>2,039</td>
<td>4,615</td>
<td>2,764</td>
<td>10,261</td>
</tr>
</tbody>
</table>
Table 5: Estimates of persistence and of the effect of class size reductions on student achievement in mathematics.

<table>
<thead>
<tr>
<th></th>
<th>Three-way unobserved heterogeneity</th>
<th>Only student unobserved heterogeneity</th>
<th>No student unobserved heterogeneity</th>
<th>No unobserved heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dynamic Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persistence</td>
<td>0.093 (0.009)</td>
<td>0.107 (0.008)</td>
<td>0.787 (0.008)</td>
<td>0.817 (0.001)</td>
</tr>
<tr>
<td>Class size reduction from thirty to:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>twenty-five</td>
<td>0.356 (0.177)</td>
<td>0.067 (0.150)</td>
<td>0.253 (0.150)</td>
<td>−0.253 (0.143)</td>
</tr>
<tr>
<td>twenty</td>
<td>0.677 (0.171)</td>
<td>0.416 (0.141)</td>
<td>0.411 (0.141)</td>
<td>−0.322 (0.131)</td>
</tr>
<tr>
<td>fifteen</td>
<td>0.814 (0.197)</td>
<td>0.409 (0.161)</td>
<td>0.466 (0.161)</td>
<td>−0.239 (0.150)</td>
</tr>
<tr>
<td>ten</td>
<td>0.398 (0.231)</td>
<td>0.020 (0.183)</td>
<td>0.176 (0.183)</td>
<td>−0.695 (0.172)</td>
</tr>
<tr>
<td><strong>Model in Gains</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class size reduction from thirty to:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>twenty-five</td>
<td>0.585 (0.295)</td>
<td>0.206 (0.225)</td>
<td>0.492 (0.150)</td>
<td>0.107 (0.138)</td>
</tr>
<tr>
<td>twenty</td>
<td>1.024 (0.274)</td>
<td>0.694 (0.214)</td>
<td>0.831 (0.140)</td>
<td>0.450 (0.126)</td>
</tr>
<tr>
<td>fifteen</td>
<td>1.268 (0.329)</td>
<td>0.831 (0.245)</td>
<td>1.043 (0.162)</td>
<td>0.675 (0.145)</td>
</tr>
<tr>
<td>ten</td>
<td>0.814 (0.388)</td>
<td>−0.080 (0.279)</td>
<td>0.483 (0.191)</td>
<td>−0.163 (0.169)</td>
</tr>
<tr>
<td><strong>Static Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class size reduction from thirty to:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>twenty-five</td>
<td>0.332 (0.171)</td>
<td>0.050 (0.145)</td>
<td>−0.634 (0.277)</td>
<td>−1.859 (0.338)</td>
</tr>
<tr>
<td>twenty</td>
<td>0.641 (0.167)</td>
<td>0.382 (0.136)</td>
<td>−1.140 (0.259)</td>
<td>−3.764 (0.314)</td>
</tr>
<tr>
<td>fifteen</td>
<td>0.768 (0.190)</td>
<td>0.358 (0.155)</td>
<td>−1.667 (0.305)</td>
<td>−4.319 (0.360)</td>
</tr>
<tr>
<td>ten</td>
<td>0.355 (0.224)</td>
<td>0.032 (0.176)</td>
<td>−0.957 (0.349)</td>
<td>−3.070 (0.405)</td>
</tr>
</tbody>
</table>

Numbers between parenthesis are standard errors. Standard errors for the first estimator are calculated as discussed in the paper. Standard errors for the rest of the estimators are two-way clustered standard errors by student and teacher. Note that it is likely that there is no asymptotic justification for using the third estimator (teacher and school unobserved heterogeneity treated as parameters that can be estimated consistently) jointly with two-way clustered standard errors. Units are the original test score units as described in the text.
Table 6: Estimates of persistence and of the effect of class size reductions on student achievement in reading.

<table>
<thead>
<tr>
<th></th>
<th>Three-way unobserved heterogeneity</th>
<th>Only student unobserved heterogeneity</th>
<th>No student unobserved heterogeneity</th>
<th>No unobserved heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dynamic Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persistence</td>
<td>0.222 (0.006)</td>
<td>0.256 (0.005)</td>
<td>0.689 (0.005)</td>
<td>0.718 (0.001)</td>
</tr>
<tr>
<td>Class size reduction from thirty to:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>twenty-five</td>
<td>0.083 (0.169)</td>
<td>-0.091 (0.118)</td>
<td>-0.133 (0.118)</td>
<td>-0.439 (0.107)</td>
</tr>
<tr>
<td>twenty</td>
<td>0.240 (0.166)</td>
<td>0.109 (0.107)</td>
<td>-0.062 (0.107)</td>
<td>-0.684 (0.097)</td>
</tr>
<tr>
<td>fifteen</td>
<td>0.270 (0.188)</td>
<td>0.166 (0.124)</td>
<td>-0.075 (0.124)</td>
<td>-0.688 (0.110)</td>
</tr>
<tr>
<td>ten</td>
<td>0.327 (0.208)</td>
<td>0.029 (0.135)</td>
<td>-0.038 (0.135)</td>
<td>-0.654 (0.118)</td>
</tr>
<tr>
<td><strong>Model in Gains</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class size reduction from thirty to:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>twenty-five</td>
<td>0.061 (0.251)</td>
<td>-0.223 (0.186)</td>
<td>0.138 (0.136)</td>
<td>-0.022 (0.107)</td>
</tr>
<tr>
<td>twenty</td>
<td>0.251 (0.249)</td>
<td>0.005 (0.169)</td>
<td>0.448 (0.127)</td>
<td>0.463 (0.096)</td>
</tr>
<tr>
<td>fifteen</td>
<td>0.305 (0.278)</td>
<td>0.132 (0.197)</td>
<td>0.573 (0.145)</td>
<td>0.621 (0.111)</td>
</tr>
<tr>
<td>ten</td>
<td>0.342 (0.325)</td>
<td>-0.111 (0.213)</td>
<td>0.445 (0.161)</td>
<td>0.159 (0.123)</td>
</tr>
<tr>
<td><strong>Static Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class size reduction from thirty to:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>twenty-five</td>
<td>0.089 (0.160)</td>
<td>-0.045 (0.105)</td>
<td>-0.734 (0.266)</td>
<td>-1.500 (0.299)</td>
</tr>
<tr>
<td>twenty</td>
<td>0.237 (0.157)</td>
<td>0.144 (0.096)</td>
<td>-1.191 (0.248)</td>
<td>-3.600 (0.277)</td>
</tr>
<tr>
<td>fifteen</td>
<td>0.259 (0.179)</td>
<td>0.178 (0.109)</td>
<td>-1.510 (0.278)</td>
<td>-4.018 (0.309)</td>
</tr>
<tr>
<td>ten</td>
<td>0.322 (0.192)</td>
<td>0.077 (0.120)</td>
<td>-1.109 (0.289)</td>
<td>-2.721 (0.324)</td>
</tr>
</tbody>
</table>

Numbers between parenthesis are standard errors. Standard errors for the first estimator are calculated as discussed in the paper. Standard errors for the rest of the estimators are two-way clustered standard errors by student and teacher. Note that it is likely that there is no asymptotic justification for using the third estimator (teacher and school unobserved heterogeneity treated as parameters that can be estimated consistently) jointly with two-way clustered standard errors. Units are the original test score units as described in the text.
Figure 1: Average test score by class size.

(a) Mathematics. 11,603 nodes out of 11,696, diameter=18, radius=11, average degree=6.985, characteristic path length=6.736.

(b) Reading. 11,949 nodes out of 12,057, diameter=18, radius=10, average degree=6.965, characteristic path length=6.787.

Figure 2: Largest connected subcomponent of $G_n$ in the estimation sample (teachers as nodes, students as edges).
Figure 3: Effect of reducing class size from thirty with pointwise confidence intervals.

Figure 4: Effect of reducing class size from thirty, estimators for dynamic model with different sources of unobserved heterogeneity.

Figure 5: Effect of reducing class size from thirty, estimators for model with all sources of unobserved heterogeneity and different restrictions on persistence.