Differential Test Performance and Peer Effects

Guido M. Kuersteiner[*], Ingmar R. Prucha[†], Ying Zeng[‡]

February 2024 - Preliminary Draft

Abstract

We use variation of test scores measuring closely related skills to isolate peer effects. The intuition for our identification strategy is that the difference in closely related scores eliminates factors common to the performance in either test while retaining idiosyncratic test specific variation. Common factors include unobserved teacher and group effects as well as test invariant ability and factors relevant for peer group formation. Peer effects work through idiosyncratic shocks which have the interpretation of individual and test specific ability or effort. We use education production functions as well as restrictions on the information content of unobserved test taking ability to formalize our approach. An important implication of our identifying assumptions is that we do not need to rely on randomized group assignment. We show that our model restrictions are sufficient for the formulation of linear and quadratic moment conditions that identify the peer effects parameter of interest. We use Project STAR data to empirically measure peer effects in Kindergarten through Third Grade classes. We find evidence of highly significant peer effects with magnitudes that are at the lower end of the range of estimates found in the literature.

[*]Department of Economics, University of Maryland, College Park, MD 20742, e-mail: gkuerste@umd.edu
[†]Department of Economics, University of Maryland, College Park, MD 20742, e-mail: prucha@umd.edu
[‡]Wang Yanan Institute for Economic Studies, Xiamen University, zengying17@xmu.edu.cn

1 Introduction

We develop a framework to analyze, identify and estimate the causal effect of peer groups on performance measures for individuals who are allocated to groups. A leading example are test scores of students who are allocated to class rooms. However, the theory could equally be applied to outcome measures of members of sports teams, teams of workers or other groups. Identification is based on the availability of multiple performance measures of comparable quality or information content. We allow team formation to be endogenous but also cover the case of randomly selected groups. For expositional purposes we refer to groups as class rooms and individuals selected into groups as students with the understanding that our procedure applies to a broader set of scenarios. For the statistical implementation we impose linear functional form restrictions. This allows us to identify the marginal effect of observed and unobserved average peer characteristics.

The causal effect of peer group composition is of interest in policy settings where overall performance of a set of individuals may be enhanced by forming groups with specific characteristics, see Lazear (2001) for a theoretical model of the benefits of higher quality peers and class size in the education context, Whitmore (2005) for an empirical analysis of the effects of class size and the share of girls per class on student achievement using Project STAR data, Carrell, Sacerdote, and West (2013) for an assessment of the effects of peer groups sorted by prior ability measures on outcomes of lowest ability students using data from the United States Air Force Academy, or Booij, Leuven, and Oosterbeek (2017) for the connection between peer group ability, tracking and outcomes using data of undergraduate economics students at the University of Amsterdam. Studies of peer group composition are part of a larger literature examining the existence and importance of peer effects.[1]

Our analysis centers around the idea of measuring the performance of individuals in closely related tasks and within a short period of time. An example are aptitude tests administered at the end of the school year and in related areas such as reading, writing and word comprehension. By considering the individual quasi-difference in scores we are able to eliminate unobserved ability and unobserved group effects. Our key identifying assumption postulates that different tests measure similar skills and that variation in differential scores is, apart from variation induced by observed covariates, due to cross-sectional variation in ability or effort of individuals that is not related to prior performance and other systematic and possibly unobserved factors correlated with group formation such as teacher quality or parental support. Our theory builds on the concept of latent performance, ability or effort measures defined in the absence of group interaction.

The first result of our paper shows that our key identifying restriction implies orthogonality

[1] See for example Hoxby (2000), Sacerdote (2001),Hanushek et al. (2003), Zimmerman (2003), Angrist and Lang (2004), Cipollone and Rosolia (2007), Duflo, Dupas, and Kremer (2011), Carrell, Sacerdote, and West (2013), Burke and Sass (2013), Booij, Leuven, and Oosterbeek (2017), Feld and Zölitz (2016), Carrell, Hoekstra, and Kuka (2018), Garlick (2018), Griffith (2022), Wu, Zhang, and Wang (2023) and for a survey Sacerdote (2011).

conditions that are at the core of our identification strategy. These restrictions are obtained without any assumptions about random group selection or independence of individual characteristics in the population of individuals. Combined with the assumption of linear latent outcomes and mean group peer effects this leads to an empirical model that falls within the class of linear peer effects models that have been prominently studied in the literature, see for example Manski (1993); Calvó-Armengol, Patacchini, and Zenou (2009); Blume et al. (2015) and Angrist (2014) for a critique of these models. We expand on this literature by explicitly accounting for endogenous peer group selection and individual heterogeneity in unobserved test taking ability. Most of the empirical literature uses observable characteristics as well as proxies for unobserved ability to measure the quality of peer groups. Our baseline model is formulated for unobserved, at least to the analyst, peer characteristics. We use multiple performance measures for similar skills to difference out common test taking ability. We allow group selection to depend on unobserved as well as observed characteristics. Recent contributions to the econometrics literature accounting for endogenous group and network formation include Goldsmith-Pinkham and Imbens (2013), Hsieh and Lee (2016) and Griffith (2022) who use an explicit network formation model, Qu and Lee (2015), Johnsson and Moon (2019), Auerbach (2022) who use control function approaches and Kuersteiner and Prucha (2020) who use an instrumental variables approach.[2]

In our setting, reduced form regression analysis that focuses on exogenous or contextual peer effects leads to biased estimates. We propose instrumental variables estimators to overcome these problems. By exploiting moment restrictions that are implied by our identifying assumptions we are proposing methods of moments estimators that identify both exogenous as well as endogenous peer effects. Formally, our moments based estimator is a panel type estimator based on quasi differences between test performance measures. An important difference between our approach and typical panel settings is the lack of a temporal dimension, and where the setup is geared towards the utilization of different test performance measures that are obtained in an essentially simultaneous fashion.

We apply our approach to the Project STAR data set of the Tennessee class size experiment. We exploit variation on closely related test score outcomes for kindergarten to third grade students. Our empirical analysis focuses on the identification, estimation and statistical inference for the parameter determining the marginal effect of unobserved peer quality. More specifically, this parameter measures the causal impact of a unit increase in peer quality, measured in terms of latent peer outcomes, on individual outcomes. Obtaining data with convincing exogenous peer

[2]The literature on peer effects is part of a larger literature on network effects. It is well recognized that the adjacency matrix frequently used in modeling peer effect is a special case of the weight matrices used in a class of models introduced by Cliff and Ord (1973, 1981), which were originally intended for modeling spatial network effects; see Anselin (2010) for a review of important contribution of this literature on identification and estimation of network effects, and Kuersteiner and Prucha (2020) for a recent contribution that connects those strands of network literature.
.

group variation is generally difficult and may require costly experimental designs. Yet, most of the empirical literature estimating peer group composition effects relies on conditionally randomly assigned peers. An advantage of the approach proposed in this paper is that we do not rely on random group assignment. Our framework enables us to identify the marginal parameter without variation in peer group composition. The random assignment assumption has been criticized for Project STAR data, for example because of attrition in higher grades, see Hanushek (2003). Our approach is designed to work despite these data limitations. In addition, and unlike in related studies identifying similar parameters such as Kelejian and Prucha (2002), Lee (2007), Graham (2008) or Kuersteiner, Prucha, and Zeng (2023) we do not rely on group size variation or variation in group type variances.

The difficulties of identifying peer effects in models with transitive groups are well known since Manski (1993). Approaches to overcome identification challenges have focused on random group assignment, for example Sacerdote (2001); Sojourner (2013); Duflo and Saez (2003); Carrell, Sacerdote, and West (2013), exogenous variation in group size as in Lee (2007) or Kuersteiner, Prucha, and Zeng (2023), instrumentation of peer effects using non-transitive group structures as in Bramoullé, Djebbari, and Fortin (2009), explicit group formation models in Goldsmith-Pinkham and Imbens (2013), variation in group specific heteroskedasticity in Graham (2008), and panel methods in Mas and Moretti (2009), Arcidiacono et al. (2012), Cornelissen, Dustmann, and Schönberg (2017) and Miraldo, Propper, and Rose (2021). The approach pursued in this paper builds on the dynamic panel methods for spatial and social networks developed in Kuersteiner and Prucha (2020).

We use a baseline model of peer effects that is similar to Arcidiacono et al. (2012) who consider peer effects working through unobserved individual characteristics. Arcidiacono et al. (2012) allow for individual specific effects as well as for endogenous group formation and use panel data. Technically, our estimator is different from theirs since we use a quasi differencing approach to eliminate class and individual fixed effects while they use a non-linear fixed effects estimator. The more important difference lies in the assumptions and empirical implementation of the method we have in mind. While our estimator could be applied to more conventional panel data with performance measures observed at different points in time, we emphasize applications where measures of closely related skills are observed essentially simultaneously. For one, this framework alleviates problems with sample attrition which are well documented for Project STAR. More importantly, our core assumption on unobserved skill or effort centers around a lack of predictive power of additional test results for closely related test measures. For example, a student scoring well on a reading test is likely to score well on a word comprehension test given at the same time. An advantage of our framework is that its credibility can be assessed and influenced by a testing protocol that complies with its basic premise. This is in contrast to more detailed restrictions between various unobserved components as well as on the evolution of individual effects measuring the change in skills over time required in the framework of Arcidiacono et al. (2012). An additional advantage of our procedure is that it does not require estimating unobserved skill or effort measures of individuals.

There is an extensive empirical literature studying educational outcomes using Project STAR data. Peer effects specifically were considered by numerous authors. Boozer and Cacciola (2001) use within and between class variation as well as controls for whether students currently are or previously were in a small classroom to estimate the parameter for endogenous peer effects. Variation in the exposure of class mates to being previously in small class rooms provides variation in peer quality that is used to identify the endogenous peer effect. They find large peer effects in second and third grade and negative but insignificant effects in first grade. Whitmore (2005) studies the effects of variation in the ratio of girls on test scores and finds mixed results depending on grades. A decomposition of the effect into endogenous and exogenous peer effects results in an estimated increase in own test score of .6 points for every point increase in the average peer test score. Graham (2008) uses a model similar to ours and estimates endogenous peer effects based on differences of the between and within variances in small and regular size classes. He finds stronger peer effects of being randomly assigned to smaller as compared to larger classes. Kuersteiner, Prucha, and Zeng (2023) interpret the variance approach of Graham (2008) as being part of a class of more general random group effects models, see also Rose (2017) for an approach similar to Graham (2008) using variance restrictions. Chetty et al. (2011) link Project STAR data with 1996-2008 tax records to investigate class room, teacher and peer effects on future earnings. Their empirical approach relies on random assignment to Kindergarten and elementary school grades which they test using additional individual level data obtained from tax records. Using analysis of variance and regression based methods exploiting within classroom variation they find significant effects of class quality on earnings. In work that looks at a related question Bietenbeck (2020) investigates the link between low ability repeaters in kindergarten classes on educational outcomes later in life and finds positive effects of being exposed to repeaters on the probability to graduate from high school and taking a college entrance exam. Pereda-Fernández (2017) relies on a conditional double randomization assumption resulting in conditional independence restrictions between student and teacher effects. These restrictions imply covariance as well as higher order restrictions that are exploited to estimate both the endogenous peer effect as well as the distribution of teacher effects. The paper finds sizable peer effects and investigates optimal teacher and class size allocation rules. Lewbel, Qu, and Tang (2023) also estimate a linear peer effects model for Project STAR data, but assuming that actual interaction between students is unobserved. They use restrictions on reduced form parameters to recover the endogenous peer effect. Empirically, they find large peer effects for second and third grade math scores.

In our empirical work we use Project STAR data and focus on classrooms in Kindergarten to Third Grade. We consider four individual SAT scores, mathematics, reading, listening and word study skills. We document the high correlation between pairs of these scores, unconditionally as well as conditionally on a full set of controls, with correlation coefficients ranging from .5 to .9. The highest correlation is observed for the reading and word study skill scores. The correlation patterns are robust across the four grade levels we consider. The peer effects coefficient of our model is estimated precisely and stable across the different grades with the exception of first grade where we observe a lower degree of peer effects. We find stronger peer effects for test score pairs that are less highly correlated, in line with our model assumptions where more closely related scores offer better control for unobserved confounding factors. Quantitatively, we estimate that an increase of average unobserved peer quality measured in SAT equivalent scores of 100 SAT points leads to an increase of 20 to 40 SAT points for an individual class mate. This estimate is based on the reading-word score pair. For other score pairs the effects range from a 40 to a 70 SAT point increase for the scores of individual class mates due to a change in average scores. These measurements are quite robust to the inclusion or omission of additional controls. When we measure exogenous peer effects, in other words the effects of average age, race and gender, as well as an indicator of economic background, we find mostly insignificant results on differential scores. This lack of significance is further evidence that our differencing strategy eliminates much of the systematic factors explaining test score variation. However, it does not mean that exogenous controls have no effect on the level of individual scores.

The paper is organized as follows. In Section 2 we develop the identification and estimation strategy. Section 2.1 defines performance measures and group selection, Section 2.2 introduces the education production function and peer effects, Section 2.3 discusses the key identifying restriction and Section 2.4 introduces the estimators we propose. Section 3 contains the empirical analysis of Project STAR data and Section 4 develops formal identification and consistency proofs as well as an asymptotic distribution theory. Section 5 contains conclusions. Proofs and tables are contained in the appendix.

2 Model

We assume that for each student $i = 1, ..., n$ we observe results on two different tests, or more generally that we observe two measures of academic achievement, say, y_{it} where $t = 1, 2$ indexes the test. Our empirical work uses Project STAR data where these measures are scores of the Stanford Achievement Tests (SAT) chosen from math, word study skills, listening and reading tests given to kindergarten through third grade students. Our approach uses two outcome measures to control for unobserved and possibly correlated student characteristics as well as group level effects related to classrooms, teachers and schools.

Let $\tilde{y}^*_{it}$ be the unobserved ability or effort of student i taking test t. We use $*$ to denote variables that are not observed by the analyst, but may be observed by the individuals, and we use a surmount $\tilde{}$ to denote variables in the original sample before group assignment. The notation $\tilde{y}^*_{it}$ emphasizes that the variable refers to the i-th student in the population of students prior to group allocation, and that the index i is an individual student identifier which is not directly tied to class rooms, e.g., i may correspond to an alphabetical ordering of all students by name . There are two steps that link the vector of latent outcomes$(\tilde{y}^*_{1t}, ..., \tilde{y}^*_{nt})$ to a vector of observed performance $(y_{1t}, ..., y_{nt})$. The first step consists in allocating student i in the original sample to classroom c where student i now receives the in-class identifier r. For example, r could be the alphabetical rank of student i in class c. Let n_c denote the size of classroom c and suppose there are C classrooms so that $n = n_1 + ... + n_C$. We then map the r-th student in classroom c into a new index, say, $i' = n_1 + ... + n_{c-1} + r$. In the following we will, abusing notation slightly, refer, e.g., to the unobserved ability of the r-th student in classroom c taking test t either as y^*_{crt} or $y^*_{i't}$, depending on the context. We will furthermore use i rather than i' for the new index for convenience of notation. The second step is an education production function or a more general group interaction model ψ that relates latent ability y^*_{crt} of classmates to actual test performance y_{crt}. Actual performance depends on class level unobserved effects, own latent ability, observed individual, class and school level covariates, as well as potential peer interaction. Our goal is to isolate the portion of observed performance that is due to peer effects.

2.1 Latent Performance and Group Selection

Selection into classrooms can be correlated with a latent baseline performance measure $\tilde{\zeta}^*_i$ that can also be thought of as prior test taking ability or latent performance, or effort prior to taking the observed tests. The purpose of introducing baseline test taking ability is to account for unobserved student characteristics that are invariant for the two tests. We account for observed student characteristics where $\tilde{v}^p_i$ and $\tilde{w}^p_{it}$ denote row vectors of observed test invariant and test varying characteristics. For notational convenience we collect the performance measures and characteristics for all n students in the sample in the following vectors and matrices of row dimension n: $\tilde{y}^*_t = (\tilde{y}^*_{1t}, \ldots, \tilde{y}^*_{nt})'$ for $t = 0, 1, 2$, $\tilde{v}^p = \left(\tilde{v}^{p'}_1, \ldots, \tilde{v}^{p'}_n\right)'$, $\tilde{w}^p_t = \left(\tilde{w}^{p'}_{1t}, \ldots, \tilde{w}^{p'}_{nt}\right)'$ for $t = 1, 2$, and $\tilde{X}^p = (\tilde{v}^p, \tilde{w}^p_1, \tilde{w}^p_2)$. The i-th row of $\tilde{X}^p$ then contains the covariates of student i in the original sample. Selection into class rooms can, in addition to depending on $\tilde{\zeta}^*$ and $\tilde{X}^p$, depend on observed school, classroom, teacher and test characteristics as well as on a vector $\alpha = (\alpha_1, ..., \alpha_C)'$ of unobserved classroom, teacher and school characteristics. We do not impose any restrictions on the cross-sectional dependence in $\tilde{y}^*_0$, $\tilde{X}$, α or observed school, classroom, teacher and test characteristics. Latent baseline performance of student i may be correlated with latent performance of student j for a variety of reasons including similar educational, socioeconomic, cultural, religious or geographic backgrounds that may or may not be known to the analyst.

We now describe the selection of students to classrooms. This assignment process results in a reordering of the students characterized by a one-to-one mapping of the original index set to the new index set. Let $\tilde{y}_t^* = (\tilde{y}_{1t}^*, ..., \tilde{y}_{nt}^*)'$ denote the vector of unobserved ability of students corresponding to the the original sample or population, let S_c be the $n_c \times n$ selection matrix that allocates students to classroom c, and let $S = [S_1', ..., S_C']'$ be the selector matrix for all students. The set of all possible class room allocations is denoted by $\mathscr{S}$. A random assignment mechanism can be thought of as selecting one element of $\mathscr{S}$ at random. Then the vector $y_t^* = (y_{1t}^*, \ldots, y_{nt}^*)$ representing the sample or population of unobserved ability for test t of students ordered by classrooms is given by

$$y_t^* = S\tilde{y}_t^*.$$

As remarked above, in slight but obvious abuse of notation, let $i = n_1 + ... + n_{c-1} + r$. Then we denote the i-th student's unobserved ability for test t interchangeably by either y_{it}^* or y_{crt}^*. The latter indexing convention, which we also employ analogously for other variables, is convenient when an analysis at the class room level is required. Now let $y_{ct}^* = \left(y_{c1t}^*, ..., y_{cn_ct}^*\right)'$, then

$$y_{ct}^* = S_c\tilde{y}_t^*.$$

Similarly, we arrange unobserved and observed student characteristics by classroom through the transformation $y_0^* = \left(y_{0,1}^*, \ldots, y_{0,n}^*\right) = S\tilde{y}_0^*$, $v^p = \left(v_1^{p'}, \ldots, v_n^{p'}\right)' = S\tilde{v}^p$, $w_t^p = \left(w_{1t}^{p'}, \ldots, w_{nt}^{p'}\right)' = S\tilde{w}_t^p$, and for $t = 1, 2$, and $X^p = \left(x_1^{p'}, \ldots, x_n^{p'}\right)' = S\tilde{X}^p = (v^p, w_1^p, w_2^p)$. Thus, the row vector x_i^p now denotes observable characteristics for student i in the sample ordered by class rooms.

We allow for S_c, $\tilde{y}_t^*$, for $t = 0, 1, 2$, and α to be mutually correlated and to be correlated with $\tilde{X}^p$ as well as with observed school, classroom, teacher and test characteristics. This assumption includes scenarios where S_c is selected completely at random, scenarios where S_c is selected based on student, teacher and school characteristics, as well as scenarios where students respond to the allocation S_c. Our assumptions, formally spelled out in Section 2.3, require that selection happens before testing which is a mild restriction that should be satisfied in most scenarios we have in mind. We do not otherwise specify or restrict the mechanism S that selects students into class rooms. We also do not require that S is observed. This is relevant in situations where we do not have information identifying individual students such as through their names, social security numbers or residential address. On the other hand we do require that we know who, among all test takers, is allocated to the same class room. For project STAR data, this information is available for the majority of students in grades K through three through a class identifier, but not for students in higher grades.

2.2 Peer Effects

We relate educational outputs to observed and unobserved inputs via an educational production function denoted by ψ. Educational production functions were considered by Krueger (1999) and

Todd and Wolpin (2003) among others. Our specification of the educational production function models observed performance y_{crt} of student r in classroom c as a function of latent ability or effort $y^*_{c1t}, \ldots, y^*_{cn_ct}$ of all students in class c. The assumptions we impose on y^*_{crt} allow it to result from a class assignment process that depends on both observed and unobserved student, teacher and school characteristics. Specifications of ψ that explicitly account for peer effects are discussed for example in Calvó-Armengol, Patacchini, and Zenou (2009), Blume et al. (2015) or Pereda-Fernández (2017).

We interpret y^*_{crt} as the unobserved test taking ability in the absence of peer effects, but accounting for possible selection of student r into classroom c, taking test t. The function $\psi_{crt}(.)$ accounts for the fact that test performance varies by student and thus depends on r as well as additional student, teacher, classroom and school characteristics not captured by y^*_{crt}. Observed performance is determined by $y_{crt} = \psi_{crt}\left(y^*_{c1t}, \ldots, y^*_{cn_ct}\right)$. The formulation of $\psi_{crt}(.)$ explicitly accounts for the possibility of peer effects by allowing for individual performance to depend on the characteristics of all peers. In the absence of peer effects the function $\psi_{crt}(.)$ simplifies to $y_{crt} = \psi_{crt}(y^*_{crt})$. In other words, the performance of the r-th student in class c only depends on own characteristics and class characteristics such as teachers and resources but not on the characteristics of other students in the class.

In empirical applications it may be difficult to work with a non-parametric framework accounting for peer effects. For this reason we focus on linear production functions with linear peer effects. Let v^c and $w^c_t, t = 1, 2$, be matrices with row dimension n that capture observed test invariant and test varying classroom, teacher or school characteristics not related to student characteristics, such as the class size, the gender, education and experience of the teacher for subject t, and let $X = (v^p, w^p_1, w^p_2, v^c, w^c_1, w^c_2)$ denote the matrix of observations on all covariates. At the class level let $y_{ct} = (y_{c1t}, \ldots, y_{cn_ct})$, $y^*_{ct} = \left(y^*_{c1t}, \ldots, y^*_{cn_ct}\right)'$, $v^p_c = \left(v^{p'}_{c1}, ..., v^{p'}_{cn_c}\right)'$, $w^p_{ct} = \left(w^{p'}_{c1t}, ..., w^{p'}_{cn_ct}\right)'$ with similar definitions for v^c_c and w^c_{ct}. We impose the following restriction on the education production function.

Assumption 1. *Let* $X^c_{ct} = (v^c_c, w^c_{ct})$, $\beta^c_t = \left(\beta^{c'}_{vt}, \beta^{c'}_{wt}\right)'$, $X^p_{ct} = (v^p_c, w^p_{ct})$, $\beta^p_t = \left(\beta^{p'}_{vt}, \beta^{p'}_{wt}\right)'$. *Assume that*

$$y_{ct} = \psi_{ct}(y^*_{ct}) = \alpha_c f_t \mathbf{1}_c + X^c_{ct}\beta^c_t + (I_c + \rho M_c)(X^p_{ct}\beta^p_t + y^*_{ct}) \tag{1}$$

where $\mathbf{1}_c = (1, ..., 1)'$ *is an* $n_c \times 1$ *vector and* M_c *is some matrix of dimension* $n_c \times n_c$ *with zero diagonal elements and* ρ *is a fixed parameter. The parameter* f_t *captures test specific interactive effects.*

The part of the expression involving M_c in (1) models peer effects. The vector $M_c y^*_{ct}$ contains weighted averages of the latent performance of class room peers, and the parameter ρ specifies the degree to which peer effects influence actual outcomes. [3] The vector $M_c X^p_{ct}\beta^p_t$ represents contextual

[3] In the usual terminology of Cliff-Ord (1973,1981) models, $M_c y^*_{ct}$ is called a spatial lag and ρ the corresponding parameter.

peer effects.

Linear specifications such as the one in Assumption 1 have been motivated as solutions to Nash games in the literature, see Calvó-Armengol, Patacchini, and Zenou (2009), Blume et al. (2015) or Pereda-Fernández (2017). Pereda-Fernández (2017) considers a full information game of student and teacher effort where student and teacher quality is known to the players and that leads to an equation like (1). Thus, Assumption 1 can be justified by assuming that student ability y^*_{ct} is unobserved by the analyst but known to class mates. Calvó-Armengol, Patacchini, and Zenou (2009) consider a network game in effort under full information where agents have linear quadratic utility. Therefore the educational production function (1) can be understood as an optimal response function where y^*_{ct} now has the interpretation of effort, unobserved by the analyst but known to peers. Finally, Blume et al. (2015) analyze a Bayes-Nash equilibrium of agents choosing effort and facing a linear quadratic utility function that depends on their own as well as their peer's effort; cp. also Cohen-Cole, Liu, and Zenou (2018) and Drukker, Egger, and Prucha (2022). Agents maximize expected utility conditional on publicly observable characteristics as well as private information that is only known to the individual player but not the other agents. They argue, see Blume et al. (2015) p. 452, that their model is observationally equivalent to Calvó-Armengol, Patacchini, and Zenou (2009) as long as the row sums of M_c are all equal, an assumption we impose below. In particular, as noted by Blume et al. (2015) p. 458, y^*_{ct} then can be interpreted as ability, with agents expectations absorbed into a fixed effect. This can be matched to our setting by assuming that y^*_{ct} is ability that is unobserved by the analyst.

The matrix M_c in Assumption 1 could be an arbitrary weight matrix describing peer interaction subject to certain measurability assumptions specified below. For exposition and concreteness we focus on the case where average characteristics of peers determine peer effects. Define

$$M_c = \left(\mathbf{1}_c\mathbf{1}_c' - I_c\right) / \left(n_c - 1\right)$$

where I_c is the $n_c \times n_c$ identity matrix. Then, M_c is the operator that computes the leave out average of peer characteristics. Using the functional form in Assumption 1 leads to an education production function $y_{crt} = \psi_{crt}\left(y^*_{ct}\right)$ for observed outcomes in terms of unobservables and observables in scalar notation $(t = 1, 2)$

$$y_{crt} = \alpha_c f_t + x^c_{ct}\beta^c_t + x^p_{crt}\beta^p_t + y^*_{crt} + \rho\left[(n_c - 1)^{-1}\sum_{l=1, l\neq r}^{n_c}\left(x^p_{clt}\beta^p_t + y^*_{clt}\right)\right]. \tag{2}$$

2.3 Identification

In this paper we propose an identification strategy for the parameter ρ that does not require randomization over $\mathscr{S}$ or variation in S_c. Our argument proceeds in two steps. First, we propose a moment restriction on unobserved test taking ability y_{it}^* and show that this restriction implies linear and quadratic moment restrictions for quantities we are able to observe in the data. We then show that ρ can be identified by considering a quasi differencing transformation of (2) and a GMM estimator that exploits the restrictions on the conditional mean and variance implied by our identifying assumption. The asymptotic theory for a general class of GMM estimators of this type was developed by Kuersteiner and Prucha (2020). Here, we rely on their theory to derive sharp identification results and inference procedures that are specific to this application.

We formalize the information structure of our model in the following definition. Recall that y_0^* stands for latent test performance, effort or ability prior to taking tests $t = 1, 2$. Also note that t is an arbitrary label such that $t - 1$ in the definition below has the interpretation of the other test. We define the following:

Definition 2.1. Let $y_{-it}^* = \left(y_{1t}^*, .., y_{i-1,t}^*, y_{i+1,t}^*, ..., y_{n,t}^*\right)'$ and $y_t^* = \left(y_{1t}^*, ..., y_{nt}^*\right)$. Define the sigma fields (information sets)

$$\mathcal{F}_{n,i,t} = \sigma\left(S, \alpha, X, \left\{\zeta_{j,0}^*, ..., y_{j,t-1}^*\right\}_{j=1}^n, y_{-i,t}^*\right) \text{ for } t = 1, 2,$$

and let $\mathcal{Z}_n^* = \sigma\left(S, y_0^*, \alpha, X, z\right)$ be the sigma field of all conditioning variables, where z denotes variables excluded from X that may be used as instruments, and let $\mathcal{Z}_n = \sigma\left(X, z\right) \subset \mathcal{Z}_n^*$ be the subset of observable information.

The information set $\mathcal{Z}_n^*$ consists both of observable characteristics X, class room allocations S, excluded variables z that may be used as instruments as well as of unobserved variables y_0^* and α. The following proportionality restriction on part of the conditional means of the latent outcomes is at the core of our proposed method of identifying peer effects.

Assumption 2. *Assume that*

$$E\left[y_{it}^* | \mathcal{F}_{n,i,t}\right] = \kappa_i f_t$$

where κ_i is a random variable that is invariant to t and is measurable with respect to $\mathcal{Z}_n^*$ and f_t is a fixed parameter that only varies with t. In addition, assume that the random variable y_{it}^* is either bounded, or there exists an $\eta > 0$ and a random variable y such that $|y_{it}^*| + |\kappa_i f_t| \leq y$ with $E\left[|y|^{4+\eta} | \mathcal{Z}_n^*\right] \leq K_y < \infty$ for all i and t. In addition, $E\left[|\alpha_c|^{2+\eta}\right] \leq K_\alpha < \infty$ for all c.

The assumption includes cases where $f_t = 1$ which is relevant in situations where tests t measure closely related skills. The interpretation of the condition is that conditional on a hypothetical or actual baseline of test results ζ^*, as well as information about group formation S, unobserved group characteristics α, as well as other observed characteristics X, additional test results do not change expected performance, except maybe for a common scale factor f_t accounting for systematic differences between tests affecting all test takers in the same fashion.[4] It is in this sense that tests $t = 1$ and $t = 2$ measure essentially identical skills. Note that $E\left[y_{it}^*|\mathcal{Z}_n^*\right] = \kappa_i f_t$ by iterated expectations and $\mathcal{Z}_n^* \subset \mathcal{F}_{n,i,t}$.

Relative to a baseline represented by $\mathcal{Z}_n^*$ the latent test results y_{it}^* are as good as randomly assigned. Conditional on observed covariates as well as the information in $\mathcal{Z}_n^*$, actual outcomes y_{it} which are determined by the educational production function in (1), then are only correlated in the cross-section because of peer effects, i.e. when $\rho \neq 0$.

The variable κ_i is a function of the entire cross section of latent performance measures ζ^*, of the allocation S and observed and unobserved characteristics. It is generally cross-sectionally dependent in ways we do not restrict or specify and captures such common effects as the socioeconomic background of students and classmates, their family background including parental education and support, their interaction with peers, the exposure to their teachers and resources available in their class room. It also depends on a student's own ability. The restriction we impose is that these factors do not change between tests $t = 1$ and $t = 2$, except maybe for a change in scale f_t.

Baseline performance ζ^* does not necessarily have to be observed or realized and ζ^* can alternatively be interpreted as latent ability prior to taking the actual tests $t = 1, 2$. In this sense, $\kappa_i f_t$ and $E\left[y_{it}^*|\mathcal{Z}_n^*\right]$ are also unobserved. A scenario where Assumption 2 is realistic arises when the same type of test is taken multiple times without additional training between iterations of the test, or when several tests are given that focus on related skills, as is the case in Project STAR data for grades K-3 and tests for reading, word study and and listening skills.[5] The assumption is less plausible if comparisons are attempted across different subjects in a high school or college setting, for tests given at different times during the school year, or measuring different skills in a professional or team setting. In the latter case, additional factors can be introduced if multiple measurements per skill category are available.[6]

The following theorem formally establishes a representation for y_{it}^* that decomposes y_{it}^* into its conditional mean, $E\left[y_{it}^*|\mathcal{F}_{n,i,t}\right]$, that is dependent cross-sectionally as well as across tests, and into uncorrelated idiosyncratic noise u_{it}. The theorem formalizes the definition of u_{it} .

Theorem 2.1. *Suppose Assumption 2 holds. Define*

$$u_{it} = y_{it}^* - E\left[y_{it}^*|\mathcal{F}_{n,i,t}\right] = y_{it}^* - \kappa_i f_t,$$

[4]For simplicity we assume that f_t in Assumptions 1 and 2 are the same. This restriction can be relaxed if three test outcomes are used. The factors f_t can be treated as random rather than fixed at the expense of slightly more complex asymptotic arguments needed for the analysis of statistical inference, see Kuersteiner and Prucha (2020).

[5]Another example of a test with multiple scores is the Comprehensive Testing Program (CTP) administered by the Educational Records Bureau and given to grades 1-11. The CTP test consists of several main categories such as 'Auditory Comprehension', 'Reading Comprehension' and 'Mathematics' with each category consisting of additional subcategories that each receive separate scores.

[6]The multi factor case is discussed in detail in Kuersteiner and Prucha (2020).

then by construction

$$E\left[u_{it}|\mathcal{F}_{n,i,t}\right]=0 \quad \textit{and} \quad \mathrm{Cov}\left(u_{it}u_{js}|\mathcal{F}_{n,i,t}\right)=0$$

for any $j \neq i$ *or* $t \neq s$*. Furthermore, let* $\mathcal{G}_{n,i,t}$ *be any sigma field with* $\mathcal{G}_{n,i,t} \subseteq \mathcal{F}_{n,i,t}$*, then* $E\left[u_{it}|\mathcal{G}_{n,i,t}\right]=0$ *and* $\mathrm{Cov}\left(u_{it},u_{js}|\mathcal{G}_{n,i,t}\right)=0$ *for any* $j \neq i$ *or* $t \neq s$*, and* $\mathrm{Cov}\left(u_{it},\kappa_i f_t\right)=0, \mathrm{Cov}\left(u_{it},X\right)=0$*, etc. If* $E\left[y^*_{it}-\kappa_i f_t|\mathcal{G}_{n,i,t}\right]\neq 0$ *for some* $\mathcal{G}_{n,i,t}$*, then Assumption 2 cannot hold.*

Examples for the information sets $\mathcal{G}_{n,i,t}$ in the above theorem are $\mathcal{G}_{n,i,t}=\sigma\left(S,\alpha,X,\zeta^*,\ldots,y^*_{t-1},y^{*s}_{-i,t}\right)$ where $y^{*s}_{-i,t}$ is a subvector of $y^*_{-i,t}$ or $\mathcal{G}_{n,i,t}=\mathcal{Z}^*_n$. The above theorem makes clear that the main restriction we impose on y^*_{it} is the assumption that latent test performance is not predictable conditional on observable characteristics and unobserved baseline ability. The properties $E\left[u_{it}|\mathcal{Z}^*_n\right]=0$, $\mathrm{Cov}\left(u_{it},u_{js}|\mathcal{Z}^*_n\right)=0$ for any $j \neq i$ or $t \neq s$ and $\mathrm{Cov}\left(u_{it},\kappa_i f_t\right)=0$, $\mathrm{Cov}\left(u_{it},X\right)=0$ follow directly from Assumption 2 and are not additional assumptions imposed on the distribution of y^*_{it}. In other words, Assumption 2 is necessary and sufficient for the decomposition of $y^*_{it}=\kappa_i f_t+u_{it}$ with uncorrelated errors u_{it}.

We do not attach a specific economic or behavioral interpretation to the decomposition of y^*_{it} into $\kappa_i f_t$ and u_{it}. It is merely a statistical representation for the purpose of isolating variation in test scores that is correlated within groups but invariant across tests and variation that is idiosyncratic to the individual student and test. Normalizing $f_2=1$, which is without loss of generality, leads to an immediate consequence of the decomposition: The quasi difference

$$y^*_{i1}-y^*_{i2}f_1=u_{i1}-u_{i2}f_1 \tag{3}$$

is mean zero and uncorrelated across i conditional on $\mathcal{Z}^*_n$. We exploit these two moment restrictions to formulate our GMM estimator. They are at the heart of our identification result which is formally proved in Section 4.

Our framework is in contrast with much of the econometrics literature on the identification of peer effects where assumptions about the orthogonality of u_{it} and $\kappa_i f_t$ and covariates are usually imposed directly. These unobservable quantities often are given economic interpretations. Examples of papers with such assumptions include Lee (2007) who assumes iid errors, Graham (2008) who assumes independent class room and idiosyncratic errors conditional on class type, Arcidiacono et al. (2012) who assume uncorrelated idiosyncratic errors and fixed effects that are orthogonal to idiosyncratic errors, Rose (2017) who assumes uncorrelated idiosyncratic errors, Kuersteiner, Prucha, and Zeng (2023) who assume a random effects specification and Lewbel, Qu, and Tang (2023) who assume exogeneity of group formation. Often these assumptions are justified by random group selection. In our setting we neither assume that individuals are randomly assigned to groups, nor that they are randomly selected from a population or that they have otherwise observed or unobserved characteristics that are independent in the cross-section.

Let $\kappa = (\kappa_1, \ldots, \kappa_n)'$ and $\tilde{\kappa} = (\tilde{\kappa}_1, \ldots, \tilde{\kappa}_n)' = S^{-1}\kappa$, then it is of interest to note that our theory allows for the variables $\tilde{\kappa}_i$ to be arbitrarily influential on the process that forms groups. An example of such correlation potentially arises in the Project Star data. Despite random allocation of students to class rooms and teachers to class rooms in kindergarten it is possible that parents who are unhappy with their allocation try and succeed to move their child to a different class or school, or that those parents provide additional training for their child. In such a scenario individual and classroom specific effects $\tilde{\kappa}_i$ are not independent of classroom allocations. In addition, attrition in first through third grades makes random assignment less plausible, see Hanushek (2003).

2.4 Estimation

Our estimators are using differential, or quasi differential test scores, $y_{i1} - y_{i2}f_1$. These differential measures purge performance measures from common unobserved influences captured by α_c and κ_c and shed light on the existence of peer effects. Define the unobserved class and individual effect as

$$\mu_c^* = \alpha_c \mathbf{1}_c + (I_c + \rho M_c)\,\kappa_c \tag{4}$$

such that assumptions 1, 2 and Theorem 2.1 imply the empirical specification

$$y_{ct} = \mu_c^* f_t \quad + v_c^c \beta_{vt}^c + w_{ct}^c \beta_{wt}^c + (I_c + \rho M_c)\,(v_c^p \beta_{vt}^p + w_{ct}^p \beta_{wt}^p + u_{ct}). \tag{5}$$

We normalize $f_2 = 1$ without loss of generality. By quasi differencing (5), which eliminates class room and individual fixed effects, we obtain

$$\begin{aligned} y_{c1} - y_{c2}f_1 &= v_c^c(\beta_{v1}^c - f_1\beta_{v2}^c) + w_{c1}^c\beta_{w1}^c - f_1 w_{c2}^c\beta_{w2}^c \\ &+ (I_c + \rho M_c)\,(v_c^p(\beta_{v1}^p - f_1\beta_{v2}^p) + w_{c1}^p\beta_{w1}^p - f_1 w_{c2}^p\beta_{w2}^p) + (I_c + \rho M_c)\,(u_{c1} - f_1 u_{c2}). \end{aligned} \tag{6}$$

We treat f_t as an unknown parameter to be estimated that accounts for differences in average test scores between test 1 and 2. Collecting terms with f_1 leads to an interpretation of (6) where $y_{c2} - v_c^c\beta_{v2}^c - w_{c2}^c\beta_{w2}^c - (I_c + \rho M_c)\,(v_c^p\beta_{v2}^p + w_{c2}^p\beta_{w2}^p)$ is used to control for unobserved μ_c^* similar to a control function approach. The difference to a conventional control function approach is that (5) for $t = 2$cannot be consistently estimated by least squares because unobserved components may be correlated with observed covariates, a problem that our GMM estimators address.

We next write (6) more compactly by stacking observations across class rooms. Let y_t , v^p and w_t^p be as defined earlier. In similar fashion define v^p, w_t^p and u_t, and let $M = \text{diag}_{c=1}^C (M_c)$. Let

$\underline{X} = (v^c, w_1^c, w_2^c, v^p, w_1^p, w_2^p, Mv^p, Mw_1^p, Mw_2^p)$ and define

$$\delta(f_1, \rho, \beta) = \left(\delta_v^{c'}, \beta_{w1}^{c'}, -f_1\beta_{w2}^{c'}, \delta_v^{p'}, \beta_{w1}^{p'}, -f_1\beta_{w2}^{p'}, \rho\delta_v^{p'}, \rho\beta_{w1}^{p'}, -\rho f_1\beta_{w2}^{p'}\right)', \tag{7}$$

then we can write (6) for the entire sample more compactly as

$$y_1 = f_1 y_2 + \underline{X}\delta + (I + \rho M)(u_1 - f_1 u_2). \tag{8}$$

With Assumptions 1 and 2, and standard regularity conditions on $\underline{X}$, we can identify δ, which is sufficient for the identification of (ρ, f_1). The parameters β_{w1}^c and β_{w1}^p can be directly recovered from δ, and β_{w2}^c and β_{w2}^p can be identified when $f_1 \neq 0$. As is common in pure fixed effect panel settings, the effects of test invariant covariates v, in our case $\beta_{v1}^c, \beta_{v2}^c$ and $\beta_{v1}^p, \beta_{v2}^p$ are not identified when using within or differencing estimators. In our case, except when $f_1 = 1$, we are quasi differencing the equation and allowing for heterogeneity in parameters between $t = 1$ and $t = 2$. This leads to a formulation of the model where v^c and v^p enter the quasi, or fully differenced equation. However, without further assumptions the parameters β_{v1}^c, β_{v2}^c, β_{v1}^p and β_{v2}^p are not separately identified. An example of restrictions where the parameters β_{v1}^c and β_{v2}^c are identified arises when $f_1 \neq 1$ and $\beta_{v1}^c = \beta_{v2}^c$.

In empirical applications it is sometimes sufficient to estimate β or even just δ, for example when $\underline{X}$ contains contextual peer effects. Conventional regression methods applied directly to (5) are invalid because unobserved fixed effects collected in μ_c^* may be correlated with $\underline{X}$. We show in Section 4 that linear instrumental variables estimation of (8) instrumenting for y_2 identifies the parameters δ and f_1. We show in Corollary 4.1 that when test scores are recorded as non-negative valuesthe class room fixed effect or an overall constant is a valid instrument. More generally, exogenous covariates that vary at the class room level but are excluded from X are valid instruments by Lemma 4.1.

Valid standard errors are obtained by clustering by class rooms. For estimation of the full set of parameters including the peer effect parameter ρ we use the efficient GMM estimator developed in Kuersteiner and Prucha (2020) for a fairly general class of networks covering both social and spatial network effects. Identification of the parameters ρ, f_1 and δ follows from Assumptions 1 and 2 as well as class room level homoskedasticity and some restrictions on the parameter space that are detailed in Section 4.

Feasible efficient inference requires some additional restrictions on conditional variances that we now describe. We allow for some forms of heteroskedasticity for u_{it} across tests and by class type. Let τ_c be a categorical variable for class type with $\tau_c = j$ if class room c is of type $j \in \{1, ..., J\}$. Examples of types are small, medium and large classes, or classes located in urban, suburban or rural areas.

Assumption 3. *Let u_{it} be as defined in Theorem 2.1 and let $i = n_1 + \ldots + n_{c-1} + r$ with $1 \leq r \leq n_c$, i.e., i refers to student r in class c. Let $u_{-i,t} = (u_{1,t}, ..., u_{i-1,t}, u_{i+1,t}, ..., u_{n,t})$. Define the filtration*

$$\mathcal{B}_{n,i,t} = \sigma\left(S, \alpha, X, \left\{y^*_{j,0}, u_{j1}..., u_{j,t-1}\right\}_{j=1}^n, u_{-i,t}\right) \text{ for } t = 1, 2.$$

Then the conditional variance of u_{it} is given by

$$E\left[u_{it}^2 | \mathcal{B}_{n,i,t}\right] = \sigma_t^2 \rho_{\tau_c}^2,$$

where τ_c is $\mathcal{Z}_n$-measurable, and $c_u \leq \sigma_t^2, \rho_j^2 \leq C_u$ for some finite positive bounds. For $t = 1, 2$, $\Sigma_t \equiv E\left[u_t u_t' | \mathcal{Z}_n^*\right] = \sigma_t^2 \operatorname{diag}_{c=1}^C \left(\rho_{\tau_c}^2 I_c\right)$.

Observing that f_t is a fixed constant, conditional on S, ζ^*, X knowledge of y^*_{it} is equivalent with knowledge of u_{it}. Thus we have the following equivalence of information sets

$$\mathcal{F}_{n,i,t} = \mathcal{B}_{n,i,t} \, \text{for} \, t = 1, 2.$$

Assumption 3 imposes a similar factor structure on conditional variances as does Assumption 2 for conditional means.

Below we illustrate the nature of our efficient quasi differencing transformation. By Theorem 2.1 the errors u_{it} are mutually uncorrelated over tests as well as cross-sectionally. Consequently the variance covariance matrix of $u_1 - f_1 u_2$ is given by,

$$\Omega = \operatorname{Var}\left(u_1 - f_1 u_2 | \mathcal{Z}_n^*\right) = \Sigma_1 + f_1^2 \Sigma_2 = \left(\sigma_1^2 + f_1^2 \sigma_2^2\right) \operatorname{diag}_{c=1}^C \left(\rho_{\tau_c}^2 I_c\right).$$

Define $\gamma_j^2 = \left(\sigma_1^2 + f_1^2 \sigma_2^2\right) \rho_j^2$, for $j = 1, ..., J$ and $\gamma = (\gamma_1^2, ..., \gamma_J^2)$. Then $\Omega(\gamma) = \operatorname{diag}_{c=1}^C \left(\gamma_{\tau_c}^2 I_c\right)$.

The disturbance term in (8) is of the form $(I + \rho M)(u_1 - f_1 u_2)$. Denote the parameter $\phi = (\rho, f_1, \delta)$. Premultiplying (8) by $\Omega(\gamma)^{-1/2}(I + \rho M)^{-1}$ then yields the following model

$$y^+(\phi, \gamma) = X^+(\phi, \gamma)\delta + u^+ \tag{9}$$

where,

$$y^+(\phi, \gamma) = \Omega(\gamma)^{-1/2}(I + \rho M)^{-1}(y_1 - f_1 y_2), \tag{10}$$

$$X^+(\phi, \gamma) = \Omega(\gamma)^{-1/2}(I + \rho M)^{-1}\underline{X}, \tag{11}$$

$$u^+ = \Omega(\gamma)^{-1/2}(u_1 - f_1 u_2) \tag{12}$$

and $\mathrm{Var}(u^+|\mathcal{Z}_n^*) = I_n$. In light of (9)-(12) we define

$$\begin{aligned} u^+(\phi,\gamma) &= y^+(\phi,\gamma) - X^+(\phi,\gamma)\delta \\ &= \Omega(\gamma)^{-1/2}(I+\rho M)^{-1}(y_1 - f_1 y_2 - \underline{X}\delta) \end{aligned} \tag{13}$$

and observe that $u^+(\phi,\gamma) = u^+$ at the true parameters.

The transformation leading to u^+ consists of three components. Quasi differencing eliminates classroom and individual effects. The operator $(I+\rho M)^{-1}$ removes cross-sectional correlation and can be understood as a form of spatial Cochrane-Orcutt transformation. Finally, the operator $\Omega(\gamma)^{-1/2}$ scales the spatially uncorrelated residuals to unit variance in the case of heteroskedasticity.

We next provide an outline of our GMM estimation methodology. For clarity we explicitly denote in the following the true parameter vectors as $\phi_0 = (\rho_0, f_{1,0}, \delta_0')'$ and $\gamma_0 = (\gamma_{1,0}, ..., \gamma_{J,0})'$. Rigorous consistency and asymptotic normality results as well as additional assumptions needed to establish these properties will be given in Section 4 below. Our GMM estimator is based on both linear and quadratic moment conditions. We consider instrument matrices H and A and require that all elements of H and A are functions of the observable variables that are measurable with respect to $\mathcal{Z}_n$. Note that by construction, these variables may depend on the unobserved effects α and κ.

In particular, we set $H = [\underline{X}, z]$ as an $n \times q$ instrument matrix.

The moment function consists of a set of linear and quadratic moments for the transformed residuals[7]

$$m_n(\phi,\gamma) = n^{-1/2}\begin{pmatrix} H'u^+(\phi,\gamma) \\ u^{+\prime}(\phi,\gamma)Au^+(\phi,\gamma) \end{pmatrix}.$$

Under the maintained assumptions of this paper the results in Kuersteiner and Prucha (2020), Theorem 1, imply that the linear moment function is uncorrelated with the quadratic moments. We have

$$E\left[m_n(\phi_0,\gamma_0)m_n(\phi_0,\gamma_0)'\right] = \frac{1}{n}\begin{pmatrix} E\left[H'H\right] & 0 \\ 0 & 2E\left[\mathrm{tr}\left(A^2\right)\right] \end{pmatrix} \equiv \Xi_n. \tag{14}$$

Therefore, our optimal weight matrix is the inverse of Ξ_n. Consistent estimates of the elements of Ξ_n can be obtained as follows. For $V_n^h = \frac{1}{n}E\left[H'H\right]$ we can use the estimator $\hat{V_n^h} = \frac{1}{n}H'H$ and for $V_n^a = \frac{1}{n}E\left[\mathrm{tr}\left(A^2\right)\right]$ we can use the estimator $\hat{V_n^a} = \frac{1}{n}\mathrm{tr}\left(A^2\right)$. The empirical criterion function of the GMM estimator then can be written as

$$Q_n(\phi,\gamma) = u^{+\prime}(\phi,\gamma)H\left(H'H\right)^{-1}H'u^+(\phi,\gamma) + \frac{\left(u^+(\phi,\gamma)'Au^+(\phi,\gamma)\right)^2}{2\,\mathrm{tr}\left(A^2\right)} \tag{15}$$

[7]The specification can be readily extended to include several quadratic moment conditions along the lines of, e.g., Kuersteiner and Prucha (2020).

such that the GMM estimator for ϕ_0 is defined as $\hat{\phi}(\gamma) = \text{argmin}_\phi Q_n(\phi, \gamma)$ for a fixed value of γ. In applications γ is replaced with a first step consistent estimator. We discuss consistent estimation of γ and feasible versions of $\hat{\phi}(\gamma)$ in Section 4. A simple two-step procedure to obtain initial estimates for δ, f_1, and ρ may sometimes be sufficient and can be used to estimate γ. The first step relies on linear moment conditions for the identification of the parameters δ and f_1, setting $\rho = 0$. The idea is to concentrate out δ and f_1 in a first stage using the linear moment function. Note that by Lemma 2.1 the linear moment condition $E[H'(y_1 - y_2 f_1 - \underline{X}\delta)] = 0$ holds at the true parameter values for δ and f_1. This equation corresponds to a regular just identified linear instrumental variables problem with included exogenous covariates $\underline{X}$ and where z is instrumenting for the endogenous variable y_2. Letting $W = (\underline{X}, y_2)$ and solving the sample analog of the moment condition yields

$$\begin{pmatrix} \tilde{f}_1 \\ \tilde{\delta} \end{pmatrix} = \left(H'W\right)^{-1} H'y_1. \tag{16}$$

The estimator for the coefficient f_1 has the familiar form of just identified two stage least squares with included covariates

$$\tilde{f}_1 = \frac{z'Q_X y_1}{z'Q_X y_2} \tag{17}$$

where Q_X is the residual operator of a projection onto $\underline{X}$. In the second step we estimate ρ using a form of the quadratic moment function, where we plug into the residual vector $\tilde{\delta}$ and $\tilde{f}_1$ for δ and f_1. Let

$$\epsilon^+(\phi) = \epsilon^+(\rho, f_1, \delta) = (I + \rho M)^{-1}(y_1 - f_1 y_2 - \underline{X}\delta)$$

and consider the quadratic moment vector

$$m_\epsilon^q(\rho) = \epsilon^+\left(\rho, \tilde{f}_1, \tilde{\delta}\right)' A \epsilon^+\left(\rho, \tilde{f}_1, \tilde{\delta}\right)$$

with A is defined as before. The first stage estimator for ρ then is obtained as

$$\tilde{\rho} = \text{argmin}_\rho \left(m_\epsilon^q(\rho)\right)^2. \tag{18}$$

Consistency and the asymptotic distribution of $\tilde{\phi}$ is discussed in Section 4 together with a discussion of the efficient estimator $\hat{\phi}$.

3 Empirical Results

In this section, we apply our identification strategy to the Tennessee's Project STAR (student-teacher achievement ratio) data, exploring peer effects among students in Kindergarten through

Grade 3. [8]Project STAR is a randomized experiment aiming at studying the impact of class size reduction on children's development. The data set has been widely used in studies of class size, peer effects, teacher effectiveness and other education-related topics (e.g., Krueger 1999; Dee 2004; Graham 2008; Chetty et al. 2011; Mueller 2013). Here we briefly discuss the details which are relevant for our study. For a complete description of the project and data, see Word et al. (1990); Mosteller (1995); Boyd-Zaharias et al. (2007).

Project STAR was carried out in Tennessee from 1985 to 1989, with additional data on the participants collected after the project ended. The experiments followed the kindergarten cohort of 1985 in participating schools for four years, from Kindergarten to the third grade. At the start of the 1985 academic year, participating schools randomly assigned kindergarten students and teachers into small classes (with an intended size of 13-17 students), regular classes (with an intended size of 22-25 students) and regular classes with a full-time teacher's aide. In actual implementation, small classes included 11 to 20 students, while regular classes (with or without aides) included 15 to 30 students. Randomness of the initial class assignment is ensured by the careful implementation and has been examined by a number of studies, e.g., Krueger (1999); Chetty et al. (2011). However, nonrandom attrition, switching and migration might have happened in higher grades (Hanushek, 1999). Our analysis uses the sample of each of the four grades, as our estimator remains consistent even without random assignment. The bottom panel of Table 1 summarizes the number of students and classes in each grade. While there are 11601 students in the experiment in total, the sample size in each grade is between 6325 and 6840 due to students migrating in or out of the participating schools. The total count of schools is 79 at the Kindergarten level, but it decreases to 76 at the Grade 1 level and subsequently to 75 at the Grade 2-3 level. This decline can be attributed to schools withdrawing from the STAR program. There are between 325 and 340 classes in each grade, with 124 to 140 of them being small classes.

We collect raw Stanford Achievement Tests (SAT) scores in mathematics (math), reading (read), listening (list) and word study skills (word) as test $t = 1$ and test $t = 2$ in our model. SAT is a nationally standardized test with scores comparable across grades. The test was administered on state specified examination days, occurring between late March and early April each year (Word et al., 1990; Krueger, 1999). The narrow time frame for the tests lends credence to the assumption that student's ability and preparation remain stable over the testing window. SAT scores for these four subjects are available from Grade K to Grade 3. The mean and standard deviations of scores for the four SAT tests are in the third panel of Table 1. The four tests generally share comparable means, except that the mean listening score is much higher than others in Grade K. Notably,

[8]We downloaded the data from Harvard Dataverse (https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/10766), which differs slightly from the version used by Graham(2008) and Rose(2017). For example, they construct class ID using teacher characteristics while our data has class ID readily available. Their kindergarten sample has 6,172 student, and ours have 6,325. We believe our dataset is more trustworthy as it matches the description in Word et al. (1990) in terms of sample size and summary statistics.

reading and word study skills scores have means that are close in all grades. With four test scores, there are 12 pairs of y_1 and y_2. We report results for only six unique combinations, as reversing the order of y_1 and y_2 does not change the estimate of ρ, transforms f_1 to its reciprocal, and and will scale the coefficients of control variables by $-f_1$.

Our main model includes a set of student and teacher characteristics, the mean and standard errors of which are summarized in the top panel of Table 1 by grade. Between 56.0% and 59.2% of students receive a free lunch during the experimental period, an indicator of low family income. To account for the influence of race, we employ an indicator for black students, as more than 98% of the minority students are of black ethnicity. The share of black students ranges from 32.6% to 34.7%. Girls make up 48.0% to 48.6% of the sample. We calculate age as of April 1st in each year, which is roughly the SAT examination date, using students' dates of birth. Teacher's characteristics are weighted by class size. Between 16.5% and 20.9% of the students are taught by black teachers. Meanwhile, 34.6% to 44.2% of the students have teachers with a master's degree or higher. Years of experience of the teachers are between 9.3 and 13.9.

In our preferred specification, we include four types of control variables: school fixed effects (excluding one dummy for the first school), class type fixed effects (excluding the dummy for small classes), student characteristics (including indicators for free lunch, black ethnicity, female gender, and age) and teacher characteristics (including indicators for black teachers, having master's degree or higher, and years of experience) and peer characteristics – specifically, the leave-out-mean of student characteristics. We start from the specification with only school fixed effects and subsequently introduce other types of control variables incrementally. We allow for heteroscedasticity of u between small classes and regular classes (with or without teacher's aide). Standard errors are clustered at the class level. The estimators for f_1 and ρ are summarized in Tables 6 and 7 respectively.

Before proceeding to discussing the main estimates of peer effects, we present several summary statistics to validate our model specifications and assumptions. Our estimator relies on the premise that tests $t = 1$ and $t = 2$ measure similar skills within a brief time frame. The combination of reading and word study skills aligns most closely with this narrative. To document the close connection between test scores, we use four measures to evaluate the correlation between various SAT scores from Grade K to Grade 3 in Table 2. The first two measures on the left panel are Spearman correlations between y_1 and y_2, and between $Q_X y_1$ and $Q_X y_2$. Here $Q_X = I - X(X'X)^{-1}X'$ denotes the residual projection matrix of X, where X encompasses the complete set of control variables: school fixed effects, class type fixed effect, student characteristics, teacher characteristics, and peer characteristics. The right panel of Table 2 reports the pseudo R^2 from 2SLS, i.e., the Spearman correlation between y_1 and $\hat{y}_1$, or between y_2 and $\hat{y}_2$. Here $\hat{y}_1$ is the predicted value of y_1 in the equation $y_1 = f_1 y_2 + X\delta + u$ estimated by 2SLS, where y_2 is instrumented by a constant term.

Similarly, $\hat{y}_2$ corresponds to the counterpart obtained by reversing the order of y_1 and y_2. [9]The different correlation measures paint a consistent picture of the correlation between the six reported pairs as well as across grades. The listening score is generally a bit less correlated with other measures. The math scores are somewhat more correlated with reading and word skills than with listening. We find the strongest connection between word study skills and reading scores which exhibit substantial correlations of approximately 0.9 across all four measures and in all grades. This finding is consistent with these scores measuring skills that are both related to language comprehension. Correlations between other pairs range between 0.5 and 0.8. While a key identifying assumption, Assumption 2 is not directly testable because it relates to unobserved ability, finding strong correlation between test scores is consistent with a setting where test scores have common conditional means. If this interpretation of the results in Table 2 is correct then we would expect the word-read pair of scores to provide the best control for unobserved individual and class effects that may be confounding measures of peer effects at the class room level. Our model implies that conditional on covariates correlation between tests only is due to unobserved interactive effects $\mu_c^* f_t$. While not a test for correct specification, our results are consistent with $\mu_c^* f_t$ accounting for a large fraction of the variation in test scores.

Further evidence of the ability of test score differences to control for unobserved class room and individual effects is the size of the estimated coefficient f_1. If scores measure closely related skills we expect f_1 to be close to one. In that scenario Assumption 2 is interpreted to imply that unobserved individual specific and class room level effects captured by μ_c^* affect both scores in the same way. We report estimated values of f_1 for the set of test score pairs in Table 6. Point estimates for f_1 are obtained by our efficient GMM estimator for all grades and for specifications with just a school effect in Column (1), school and class type effects in Column (2), with a full set of controls excluding contextual peer effects in Column (3) and the full set of controls in Column (4). The coefficient f_1 is estimated with high precision in Columns (1) through (3) and somewhat less precisely in Column (4). The point estimates are generally close to one with a few exceptions, such as for the listening scores in grade K and specifications (1)-(3) as well as the listening scores in grades two and three for specification (4). The parameter estimates for the read-word score combination are remarkably stable across all specifications and across all grades, are close to one and estimated precisely. This is further evidence that the read-word score pair is best suited to control for unobserved class level effects.

We explore one additional approach to evaluate the plausibility of Assumption 2. If unobserved individual and class effects μ_c^* are stable over time then scores from previous years should be good predictors for current year scores. Here, it is important to note that our theory does not

[9]We also adjust for missing values using a similar method as the one for the GMM estimator. That is replacing the peer averages of the observed individuals $\bar{x}_{(-r)c}^{obs} = (n_c \bar{x}_c - x_{rc}^{obs})/(n_c - 1)$ with $\tilde{x}_{(-r)c}^{obs} = (n_c \bar{x}_c^{obs} - x_{rc}^{obs})/(n_c - 1)$, where $\bar{x}_c^{obs}$ is the average of all observed x, Although this makes little difference to the results.

assume μ_c^* to be time invariant between grades. Conceptually, ζ^* measures unobserved ability just prior to taking the tests in the current year. As a result, μ_c^* similarly reflects only current year unobserved ability and other class room level effects coming from teachers and peers. Unobserved effects μ_c^* are expected to vary year by year for various reasons, including the fact that instruction during the year improves skills, that teachers may change and that students may move to different classes. However, it is expected that at least some components of μ_c^* that are more closely related to individual student ability are time invariant. We examine this hypothesis by measuring the explanatory power of lagged scores on current scores. The regression sample includes stayers in each grade, defined as individuals participating in the STAR project for both the previous and the current grades to ensure availability of both current and lagged scores. The Grade 1 sample includes 66% stayers, while Grade 2 and 3 samples have 73.8% and 79.6% respectively. Among the stayers in Grade 1, 88.7% stay in the same type of classes (small v.s. regular(w/wo) aide) in the same schools. This proportion exceeds 93% for Grade 2 and Grade 3. [10] The share is lower if we consider regular and regular/aid as different types, due to a random reallocation of students in regular and regular/aid classes to these two types within school in Grade 1. However, the literature generally finds no significant impact of teacher's aide (e.g., Krueger, 1999). The stability of school and class types among the large share of stayers fosters a steady learning environment. While newcomers and leavers might marginally influence peer quality, the allocation of newcomers to different classrooms is random within schools. Consequently, when accounting for school fixed effects, we anticipate a consistent level of peer quality over time.

The relatively stable environment that these students are in suggests that μ_c^* and f_t should contain significant time invariant components that in turn imply high correlation between observed scores and lagged scores. As a result we expect to see relatively high R^2 in predictive regressions with lagged scores as controls in addition to the observed characteristics. Consequently, we regress test outcomes on the full set of control variables (school and class type fixed effects, students, teacher and peer characteristics and a constant term), and then add lagged outcomes of the same or different type and their peer averages to assess changes in R^2. The results are summarized in Table 3. The results show the predictive power of lagged scores for the same test in bold, as well as when using other tests instead, for grades one through three. The predictive power of lagged scores is documented by comparing the R^2 with and without adding lags. For example, for the word score, adding its own lag improves the R^2 from .244 to .435 in grade one, from .25 to .539 in grade two and from .229 to .488 in grade three. Using the reading score, which is the score of the test most closely related to the word score, leads to increases in the R^2 that are even slightly higher. Overall, the coefficients on lagged scores when using the own lag of word as well as the lagged reading score are large in magnitude and highly statistically significant. These findings hold up for the other test score pairs with similar magnitudes and statistical significance. As may be expected, the predictive

[10] Based on the authors' calculation using the STAR data.

effects are slightly weaker for listening scores than for math, read and word. The results are further evidence that the data distributions are consistent with key features of our model.

Our analysis thus far has documented that pairs of tests scores are highly correlated both contemporaneously as well as over the course of one school year. This evidence is consistent with restrictions imposed by Assumption 2 and the hypothesis that closely related scores depend on the same unobserved components. A caveat is that closely related scores may remove too much common variation when quasi differences of these scores are used and thus render statistical evidence about peer effects imprecise. Table 4 reports pseudo R^2 measures of model estimates for the six pairs of scores and model specifications (1) through (4) with column (1) only controlling for school fixed effects, and column (4) being the full specification with school and class type effects, student and teacher characteristics as well as peer characteristics.

The observed pseudo R^2 measures are generally below .2 with a few exceptions for Grade K. We also see that the R^2 is a bit higher for models with listening scores. More over and as expected the pseudo R^2 generally increases with larger models, or in other words by moving from Column (1) to (4). The read-word score pair often has the lowest R^2 which is an indication that the differential score in these specifications already controls for most of the variation explained by the covariates. We take this as further evidence that read-word best fits the model assumptions we impose.

Table 5 presents standard deviations of $Q_x y$ for four test scores and $Q_x(y_1 - f_1 y_2)$ for six pairs of test scores for Grade K to 3. The left and right panels are for small and regular classes respectively. Control variables for specifications (1)-(4) are the same as those in Table 4, as discussed above. Estimates of f_1 are obtained from our efficient GMM estimator and presented in Table 6. We see that in Grade K and Grade 1, the ratio of standard deviations of $Q_x(y_1 - f_1 y_2)$ to that of $Q_x y$ is around 1/2 for most pairs of tests and 1/5 to 1/4 for the read-word pair. In Grade 2 and 3, this ratio is around 1/3 for most pairs and about 1/5 for the read-word pair. This shows that quasi-differencing removes a significant part of the variation in the test scores, especially for the read-word pair, consistent with the term $\mu^* f_1$ playing a major role in determining test performance. At the same time, the results in Table 4 indicate that the inclusion of controls does not eliminate much of the remaining variation not captured by fixed effects. The conclusion is that there is between 20 to 30 percent of test score variation remaining for the identification of ρ.

Table 7 presents estimates and standard errors (in parentheses) for peer effects ρ. To provide some context and interpretation of ρ consider the education production function in Equation (1). When $\rho = 0$ such that there are no peer effects it follows that the observed scores are $y_{ct} = \alpha_c f_1 \mathbf{1}_c + X_{ct}^c \beta^c + (X_{ct}^p \beta^p + y_{ct}^*)$ where y_{ct}^* has the interpretation of an unobserved (error) term that is measured in the same units as the test score variable. When ρ is different from zero test score performance of student i changes by ρ times the leave-out-average of y_{ct}^*, through the term $\rho M_c y_{ct}^*$. This interpretation implies that ρ measures the effect on student i's test score as a fraction of the increase in peer ability measured in units of test score performance. Similarly, the effect on test scores as a result of changes in peer characteristics are measured as $\rho\beta^p$ times the changes in relevant peer characteristics.

Standard errors are clustered at the class level. Test types of y_1 and y_2 are indicated in Columns (1) and (2). Control variables for the four specifications are outlined at the bottom of the table. Specification (1) includes only school fixed effects. Specification (2) further adds class type fixed effects. Specification (3) introduces student characteristics ($free\,lunch$, $black$, $girl$, age) and teacher characteristics ($black$, $master$, $experience$). Specification (4) further controls for peer characteristics, defined as the leave-out-mean of the four student characteristics variables.

The comparison of the four specifications demonstrates the robustness of our estimators to additional controls. We now focus on Specification (4), which includes the full set of control variables. Across all grade levels and all test pairings, our results consistently reveal positive and statistically significant peer effects, ranging from 0.2 to 0.7. The results we find are relatively stable across the four specifications in Columns (1)-(4) with the magnitude of the coefficient ρ generally decreasing somewhat for the larger models. The results for read-word are more robust in this sense especially in Grades K and one. Estimated peer effects tend to be largest for specifications involving the listening score as well as the math-word score. The results for Grades K, two and three are quite similar, while results for Grade one are generally somewhat smaller in magnitude but remain statistically very significant. Importantly, when examining the pairing of reading and word study skills—a context wherein the tests evaluate closely related abilities as documented by our prior analysis—the effects are notably smaller than for the remaining pairings, ranging from 0.2 to 0.4 across different grade levels. The implications of these results are that improvements in peer quality measured in terms of potential SAT scores y_t^* translate at the rate of 20 to 40 percent to improvements of individual scores or in other words a 100 point increase in the peer potential SAT score results in a 20 to 40 point increase in individual SAT scores. These effects are at the lower end of the spectrum of results reported in the literature.

The comparison with other results in the literature is complicated by the fact that often a model with endogenous peer effects is estimated. To focus ideas, consider the case without covariates and only one outcome measure y_1. Then, an endogenous peer effect specification is $y_1 = \lambda M y_1 + u$. If peer effects are measured with full rather than leave-one-out means this formulation is identical to our model which is $y_1 = (I + \rho M)\, u$ in this simplified stylized setting. The parameters ρ and λ satisfy the one-to-one mapping $\lambda = \rho/(1+\rho)$ which is a consequence of the fact that $(I + \rho M) = (I - \lambda M)^{-1}$. The inverse captures multiplier effects of peer performance inherent in the endogenous peer effect formulation, while ρ is a summary measure of all these multiplier effects. When peer effects are measured by leave-one-out rather than full means the relationship between the two formulations is no longer exact. However, it can be shown that $(I + \rho M_c)^{-1} = (I - \lambda M_c)$ continues to hold approximately with an error that is $O\left(n_c^{-2}\right)$ where n_c is the size of class c. Based on this approximation, we convert our measure of ρ to an endogenous peer effect $\lambda = \rho/(1+\rho)$, or equivalently, $\rho = \lambda / (1 - \lambda)$ for the prupose of the discussion that follows.

In comparison with existing results for peer effects employing STAR data, our estimates are relatively small in magnitude but estimated with high precision and statistical significance. Using normalized kindergarten SAT scores, Graham (2008) derives endogenous peer effects of 0.46 and 0.56 for math and reading respectively, which translates to $\rho = 0.86$ or $\rho = 1.30$ in our context.[11] Using normalized Kindergarten SAT math score as outcome, Rose (2017) finds an endogenous peer effect of 0.65 or a correlated effect (measured by coefficients of the peer average errors) of 0.71. When both are considered simultaneously, endogenous and correlated effects are 0.90 and -0.03, albeit with standard deviations over 1.7, indicating weak identification. Boozer and Cacciola (2001), using average percentiles of math and reading scores as outcomes and an instrumental variable methods for estimation, identify endogenous peer effects of 0.30, 0.86 and 0.92 for Grades 1 to 3 respectively, with standard deviations of 1.0, 0.12 and 0.04. Sojourner (2013) uses first-grade averages of SAT percentiles of math, reading and listening as outcome measure, and checks the exogenous peer effects of lagged peer outcomes. He finds an estimate of about 0.35 with a standard deviation around 0.14.

At the upper end of the spectrum are estimates by Lewbel, Qu, and Tang (2023) who report endogenous peer effects of 0.85 and 0.92 for small and regular classes in third-grade SAT math scores, each with standard deviations around 0.02. Translated to our parametrization, these estimates correspond to $\rho = 5.6$ and $\rho = 11.5$ respectively, using the approximation $\rho = \lambda/(1 - \lambda)$. The magnitude of their findings is remarkably high and suggests that a mere 10-point rise in ability equivalent peer SAT scores lead to an increase of individual scores of up to over 100 points.

The observed peer effects in our study exhibit smaller magnitudes compared to most of the results discussed above. It is important to note that estimates in Lewbel, Qu, and Tang (2023), Rose (2017) and Boozer and Cacciola (2001) pertain to coefficients of peer's average contemporaneous scores, resulting in a multiplier effect as explained above. In contrast, our estimated peer effects are coefficients of average peer unobserved ability. This notion is more akin to the "correlated effects" discussed in Rose (2017), who finds a value of 0.71. Meanwhile, Graham (2008)'s estimates can be translated to peer effects measured in our parametrization lead to equivalent values of $\rho = 0.86$ for mathematics and $\rho = 1.3$ for reading according to our definition. This indicates a noticeable difference in peer effects' magnitude between our study and those of Rose (2017) and Graham (2008).

Our proposed estimator leads to highly precise estimates compared to all the aforementioned estimates, except that of Lewbel, Qu, and Tang (2023). The enhancement in precision can be attributed in part to the elimination of individual fixed effects through differencing. Certain in-

[11] Graham (2008) does not directly estimate endogenous peer effects. His model is $y_{ic} = \alpha_c + (\gamma - 1)\bar{\epsilon}_c + \epsilon_{ic}$, with y_{ic} being the outcome of student i in class c and $\bar{\epsilon}_c$ is the full mean of unobserved ability ϵ_{ic} in class c. Hence $\gamma - 1$ is comparable with our estimate ρ, both are coefficients of average peer unobserved ability. Due to the full-mean specification, the equation is equivalent to $y_{ic} = \lambda \bar{y}_c + (1 - \lambda)\alpha_c + \epsilon_{ic}$, where $\lambda = 1 - \frac{1}{\gamma}$ is the endogenous peer effects.

dividual characteristics that have been recognized for their significant influence on scores are not observed in the data for Project STAR. These include parents' education, family income and most importantly unobserved IQ. Our estimator, by removing unobserved class effects and individual fixed effects, effectively eliminates residual variance associated with these characteristics which may be one explanation for the precise estimates with obtain. In addition, our GMM estimators use additional quadratic moment restrictions efficiently.

The aforementioned studies have each focused on one or two specific tests and are often limited to a particular grade. Methods that rely on random peer assignment are typically constrained in where they can be applied. In project STAR the random assignment assumption is more defensible in earlier grades(Sojourner, 2013; Graham, 2008; Boozer and Cacciola, 2001). In contrast, our analysis does not require random assignment and facilitates the estimation of peer effects across diverse test types and grade levels. Encouragingly, our findings consistently fall within a reasonably narrow range, affirming the robustness and reliability of our method.

We also provide the coefficients of peer averages of individual characteristics (age, black, free lunch, and girl) in Table 8. These four variables all pertain to test-invariant individual characteristics. The estimated coefficients can be expressed as $\rho(\beta^p_{v1} - f_1\beta^p_{v2})$, where $\rho\beta^p_{v1}$ and $\rho\beta^p_{v2}$ represent the exogenous peer effect of the variable on tests $t = 1$ and $t = 2$ respectively. Thus, these coefficients capture the differential exogenous peer effects across the two tests. We also scale the coefficients for dummies $girl$, $black$ and $free\,lunch$ by 1/100 so that the coefficients are the impact on the outcome of 1 percentage point changes in share of peers who are girl/black/free lunch receivers. The first observation is that, except for kindergarten where the presence of girls among peers positively impacts reading more than word study, none of the four variables show significant differential exogenous peer effects in the closely related word-read pair. Secondly, the share of black peers and the average peer age generally display insignificant differential exogenous peer effects. An exception is in Kindergarten, where peer age more favorably affects listening scores than reading scores. Given our general expectation of a negative impact from peers receiving free lunch and a positive impact from a higher share of girls, the results suggest that these two exogenous peer effects are stronger for reading and word study skills compared to listening and mathematical abilities in Grade 1. Similar patterns emerge in other grade levels, albeit with lower significance and consistency. Overall, the results indicate small and mostly insignificant peer effects on differential scores. This is consistent with a scenario where y_2 is essentially sufficient to control for both observed and unobserved test invariant determinants of test performance. In fact, when $f_1 = 1$ and $\beta^p_{v1} = \beta^p_{v2}$ these effects would be zero in our model and the overall evidence points to the conclusion that this is approximately the case in the sample as well.

4 Identification and Inference

4.1 Identification Conditions

We next discuss identification of our model. To establish identification we introduce additional assumptions imposed on the parameter space. We call the model identified at the parameter value ϕ_0 if the population moment conditions underlying the formulation of our estimators are only satisfied at $\phi = \phi_0$ for all $\phi \in \Phi$. This definition can be seen as an extension of classical definitions of identification such as Rothenberg (1971) for likelihood based approaches. In general, identification is neither necessary nor sufficient for consistency which is established separately, see Newey and McFadden (1994) for a discussion.

Assumption 4. *We use subscript zero to denote true parameters and treat δ_0, ignoring the restrictions given in (7), as an unrestricted parameter vector. Let $\phi_0 = [\rho_0, f_{1,0}, \delta_0]'$ with $\delta_0 = \delta(f_{1,0}, \rho_0, \beta_0)$ defined in (7). Then assume $\phi_0 \in \Phi$, where Φ is a compact subset of $\Phi_0 = (-1, 1) \times (-K_f, K_f) \times (-K_X, K_X)^{p_x}$, and where K_f and K_X are finite positive constants and p_x is the column dimension of $\underline{X}$. In addition, let $\gamma_0 = \left(\gamma_{1,0}^2, ..., \gamma_{J,0}^2\right)'$ with $\gamma_{j,0}^2 = (\sigma_{1,0}^2 + \sigma_{2,0}^2 f_{1,0}^2)\rho_{j,0}^2$. Then $\gamma_0 \in \Gamma$ with $\Gamma = [K_\gamma^{-1}, K_\gamma]^J$ where K_γ is a finite positive constant.*

Remark 4.1. The assumption implies that the parameter space for ρ is a subset of $[-K_\rho, K_\rho]$ for some positive $K_\rho < 1$.

In addition, we impose the the following assumption on A.

Assumption 5. *(i) The matrix $A = (a_{ij}) = \text{diag}_{c=1}^C(A_c)$ is measurable with $\mathcal{Z}_n$ and has zero diagonal elements. Moreover, A_c is a function of M_c and is symmetric.*

(ii) $\mathbf{1}_c' A_c \mathbf{1}_c / n_c \geq \underline{K}_a > 0$ for some constant $\underline{K}_a$.

(iii) $\sup_i \sum_{j=1}^n |a_{ij}| \leq K_a < \infty$ for some constant K_a.

Two valid choices for A are $A = M$ and $A = M'M - \text{diag}\,(M'M)$.

Assumption 6. *Group size n_c satisfies $2 \leq n_c \leq \bar{n}_c < \infty$ for some constant $\bar{n}_c$.*

Before discussing the general case, we start discussing identification of our model in the case where no additional covariates are present. Note that in the case of no covariates, the parameters δ of the covariates are omitted, $\phi = (\rho, f_1, \delta')'$ is reduced to $(\rho, f_1)'$ and (13) becomes $u^+(\phi, \gamma) = y^+(\phi, \gamma)$.

Observe that $\Omega(\gamma) = \text{diag}_{c=1}^C(\gamma_{\tau_c}^2 I_c)$ is diagonal for all admissible γ and that $\Omega_0 = \Omega(\gamma_0) = \text{Var}(u_1 - f_{1,0} u_2 | \mathcal{Z}_n)$ in light of Assumptions 2 and 3 and Theorem 2.1. Also, $u^+(\phi_0, \gamma) = \Omega(\gamma)^{-1/2}\Omega_0^{1/2} u^+$. Thus it is readily seen that for all A satisfying Assumption 5(i) we have

$$E\left[u^+(\phi_0, \gamma) | \mathcal{Z}_n\right] = 0, \tag{19}$$

$$E\left[u^{+}(\phi_0,\gamma)'Au^{+}(\phi_0,\gamma)|\mathcal{Z}_n\right] = \text{tr}\left(\Omega(\gamma)^{-1/2}A\Omega(\gamma)^{-1/2}\Omega_0\right) \tag{20}$$

$$= \text{tr}\left(A\Omega(\gamma)^{-1}\Omega_0\right) \tag{21}$$

$$= 0 \tag{22}$$

observing that the diagonal elements of $A\Omega(\gamma)^{-1}\Omega_0$ are zero.

We next discuss identification for a given matrix $H = z \equiv (z_1', ..., z_c', ..., z_C')'$ of $\mathcal{Z}_n$ measurable instruments, where the z_c are $n_c \times q$ matrices of random variables that are invariant at the class room level, i.e., $z_c = \mathbf{1}_c\dot{z}_c$, where $\mathbf{1}_c$ is an $n_c \times 1$ vector with all elements equal to one and $\dot{z}_c$ is a $1 \times q$ vector of class level characteristics. We focus on the just identified case with $q = 1$ for ease of exposition. Additional overidentifying restrictions can be added but do not affect arguments related to identification. Consider the moment vector

$$m_n(\phi,\gamma) = n^{-1/2}\begin{bmatrix} H'u^{+}(\phi,\gamma) \\ u^{+}(\phi,\gamma)'Au^{+}(\phi,\gamma) \end{bmatrix} \tag{23}$$

where A is a matrix that satisfies the properties listed in Assumption 5.

We next show that $E[m_n(\rho, f_1, \gamma)] = 0$ if and only if $(\rho, f_1) = (\rho_0, f_{1,0})$. Identification of $f_{1,0}$ can be established from an inspection of the linear moment condition. Identification of the social interaction parameter ρ_0 follows from an analysis of the quadratic moment condition.

Lemma 4.1. *As a special case of* (5) *without covariates consider the data generating process* $y_t = \mu^* f_{0,t} + (I_n + \rho_0 M)\, u_t$ *where* μ^* *is defined in* (4). *Suppose Assumptions 1, 2, 3, 5, 4, and 6 hold, let* $m_n(\phi,\gamma)$ *be defined in (23) and let* $H = z$. *In addition, assume that* $q = 1$ *and*

$$n^{-1}\left|E\left[y_2'\Omega(\gamma)^{-1/2}z\,|\mathcal{Z}_n\right]\right| \geq K_y > 0 \tag{24}$$

for all $\gamma \in \Gamma$ *and all* n, *where* K_y *is a constant. Then for all* $\gamma \in \Gamma$ *we have* $E[m_n(\rho_0, f_{1,0}, \gamma)\,|\mathcal{Z}_n] = 0$ *and* $E[m_n(\rho, f_1, \gamma)\,|\mathcal{Z}_n] \neq 0$ *a.s. for all* $(\rho, f_1) \neq (\rho_0, f_{1,0})$ *and* $\gamma \in \Gamma$.

A proof of Lemma 4.1 establishing the identifiability of $(\rho_0, f_{1,0})$ is given in the appendix. The subsequent corollary considers a special case of the lemma that is of interest in the context of test scores, and arises when $H = \mathbf{1}_n$ and when test scores are non-negative. The latter is typically the case for non-normalized test scores.

Corollary 4.1. *Suppose the assumptions of Lemma 4.1 hold,* $H = \mathbf{1}_n$, $y_t \geq 0$, *and suppose condition (24) is replaced by* $n^{-1}\sum_{c=1}^{C} E[y_{c2}'\mathbf{1}_c\,|\mathcal{Z}_n] > K_y > 0$ *for some constant* K_y. *Then the conclusions of Lemma 4.1 hold.*

We emphasize further that no restrictions on μ^* (other than the fact that μ^* does not depend on t and $\mu^* \neq 0$ a.s. and is uniformly bounded in absolute value) are needed to identify ρ_0. Also, the analysis in the appendix shows that neither knowledge of, nor restrictions on heteroskedasticity are necessary for the identification of ϕ_0. Thus, initial estimates of ϕ_0 can be obtained with $\gamma_j^2 = 1$ for all $j = 1, ..., J$.

Next we turn to a discussion of identification of $\phi = (\rho, f_1, \delta')'$ when the model contains covariates. We use instruments $H = [\underline{X}, z]$ for linear moment restrictions and a matrix A for quadratic moment restrictions, with A defined as before. Both H and A are $\mathcal{Z}_n$-measurable and $\text{diag}(A) = 0$. For ease of exposition we focus again on the case where z is a vector. Recall that with covariates, $u^+(\phi, \gamma)$ is defined in (13).

We consider the following $(q+1) \times 1$ vector of linear and quadratic moment functions:

$$m_n(\phi, \gamma) = \begin{pmatrix} m_n^{(l)}(\phi, \gamma) \\ m_n^{(q)}(\phi, \gamma) \end{pmatrix} = n^{-1/2} \begin{pmatrix} H'u^+(\phi, \gamma) \\ u^+(\phi, \gamma)'Au^+(\phi, \gamma) \end{pmatrix} \tag{25}$$

where the functions $m_n^{(l)}(\phi, \gamma)$ and $m_n^{(q)}(\phi, \gamma)$ denote the linear and quadratic moment conditions respectively.

Note that our procedure is not designed to identify γ, nor is that required for the identification of ϕ. Lemma 4.2 below shows that for all premissible γ the moment condition satisfies $E\left[m_n(\phi_0, \gamma)\,|\mathcal{Z}_n\right] = 0$ and $E\left[m_n(\phi, \gamma)\,|\mathcal{Z}_n\right] \neq 0$ for $\phi \neq \phi_0$. The parameters γ can be pinned down by estimating the variance of $u_1 - f_{1,0}u_2$ for different types of groups. For identification, we further impose the following assumption.

Assumption 7. *(i)The absolute values of the elements of h_{ir} of $H = [\underline{X}, z]$ and μ_i^* in μ^* are uniformly bounded in i, r, n by some positive constant $K_H < \infty$.*

(ii) The smallest eigenvalue of $\underline{X}'\underline{X}/n$ is bounded from below by some $\xi_X > 0$ uniformly in n.

(iii) Define $V(\rho, \gamma) = \Omega(\gamma)^{-1/2}(I + \rho M)^{-1}$ and let

$$Q_{V^{1/2}X}(\rho, \gamma) = I - V(\rho, \gamma)^{1/2}\underline{X}\left(\underline{X}'V(\rho, \gamma)\underline{X}\right)^{-1}\underline{X}'V(\rho, \gamma)^{1/2}$$

be the projection matrix onto the orthogonal complement of the column space of $V^{1/2}\underline{X}$, then

$$\inf_{\gamma \in \Gamma, \rho \in [-K_\rho, K_\rho]} n^{-1}\left|E\left[(V(\rho, \gamma)^{1/2}z)'Q_{V^{1/2}X}(\rho, \gamma)V(\rho, \gamma)^{1/2}y_2\,|\mathcal{Z}_n\right]\right| > K_y > 0 \text{ a.s..} \tag{26}$$

Remark. Note that the elements of the matrix H are by construction measurable w.r.t. $\mathcal{Z}_n = \sigma(X, z)$.

Since $\underline{X}$ contains X, part (i) of the assumption implies that also the elements of X are bounded in absolute value by K_H. Observing that $(\underline{X}, z)$ has full column rank by Assumption 7(iii), as setting $\rho = 0$ and $\gamma = (1, ..., 1)$ in $V(\rho, \gamma)$, (26) becomes $n^{-1}\left|E\left[z'Q_X y_2|\mathcal{Z}_n\right]\right| > K_y > 0$, where $Q_X = I - \underline{X}\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'$. It is readily seen from an inspection of the proof of Lemma 4.2 that a necessary condition for Assumption 7(iii) is that $(\underline{X}, E\left[\mu^*|\mathcal{Z}_n\right])$ has full column rank. The latter condition is equivalent to the condition that $(\underline{X}, E[y_2|\mathcal{Z}_n])$ has full column rank, observing that in light of (5) we have $E[y_2|\mathcal{Z}_n] = E[\mu^* + \underline{X}\beta_{2,0}|\mathcal{Z}_n]$.

Condition (26) generalizes condition (24) in Lemma 4.1. Under homoskedasticity or when estimators are not taking heteroskedasticity into account, the latter condition takes the more familiar form $n^{-1}|E[y_2'z|\mathcal{Z}_n]| > K_y > 0$. Further simplifications of (26) are obtained if in addition to heteroskedasticity, spatial correlation is ignored by setting $\rho = 0$. Considering this case is useful when discussing first step estimators for δ and f_1. For the case with covariates (26) simplifies to $n^{-1}|E[z'Q_X y_2|\mathcal{Z}_n]| > K_y > 0$, where Q_X is defined as above. Both versions of the simplified identification conditions are conventional rank conditions for linear instrumental variables estimators. A second scenario under which the conditions can be simplified is when γ is evaluated at a fixed constant level. For example, when $\gamma = \bar{\gamma}$ then (24) is replaced by $n^{-1}\left|E\left[y_2'\Omega(\bar{\gamma})^{-1/2}z\,|\mathcal{Z}_n\right]\right| > K_y > 0$ and and a similar simplification obtained for (26).

The following Lemma is a generalization of Lemma 4.1.

Lemma 4.2. *Let Assumptions 1, 2, 3, 4, 5, 6 and 7 be satisfied. Let* $m_n(\phi, \gamma)$ *be defined as in* (25)*. Then for all admissible* $\gamma \in \Gamma$, $E[m_n(\phi_0, \gamma)|\mathcal{Z}_n] = 0$ *and* $E[m_n(\phi_0, \gamma)|\mathcal{Z}_n] \neq 0$ *for* $\phi \neq \phi_0$ *a.s.*

4.2 Consistency of Parameter Estimators

Formulating the moment conditions using transformed residuals as in (25) requires consistent estimates of the variance parameters γ. These can be obtained by first obtaining consistent but inefficient parameter estimates. To do so, we set all elements of $\gamma = (\gamma_1^2, ..., \gamma_J^2)$ to one. Lemma 4.2 implies that the parameters $\phi = (\rho, f_1, \delta)$ are still identified by the moment conditions in (25).

First step estimators $\left(\tilde{\delta}, \tilde{f}_1, \tilde{\rho}\right)$ are defined in (16) and (18). To formalize the identification result for these non-efficient first step estimators define the moment vector

$$m_\epsilon(\phi) = n^{-1/2}\begin{pmatrix} H'(I + \rho M)\epsilon^+(\phi) \\ \epsilon^+(\phi)' A\epsilon^+(\phi) \end{pmatrix}. \tag{27}$$

The next lemma shows that the moment conditions in (27) are sufficient to identify the parameters $\phi = (\rho, f_1, \delta')'$.

Lemma 4.3. *Let Assumptions 1-6, and Assumption 7(i),(ii) be satisfied. Let* $m_\epsilon(\phi)$ *be defined as in* (27)*. Assume further that* $n^{-1}|E[z'Q_X y_2|\mathcal{Z}_n]| > K_y > 0$ *a.s . Then* $E\left[n^{-1/2}m_\epsilon(\phi_0)|\mathcal{Z}_n\right] = 0$ *and* $E\left[n^{-1/2}m_\epsilon(\phi)|\mathcal{Z}_n\right] \neq 0$ *if* $\phi \neq \phi_0$ *for all* $\phi \in \Phi$.

We next show that the above defined initial estimators are consistent. To establish consistency we impose assumptions related to the convergence of sample averages and moments of sample averages to well defined limits. These assumptions are adaptations of Assumptions 2, 3 and 5 in

Kuersteiner and Prucha (2020) to the present setting which differs somewhat from the framework of sequential exogeneity in that paper.

Let $i = n_1 + \ldots n_{c-1} + r$ with $1 \leq r \leq n_c$. Then, with some abuse of notation, denote the i-th diagonal element of $\Omega(\gamma) = \text{diag}_{c=1}^C \left(\gamma_{\tau_c}^2 I_c\right)$ as $\gamma_{(i)}^2 = \gamma_{\tau_c}^2$, noting that τ_c varies with i. Let the linear and quadratic moment vectors $m_n^{(l)}(\phi, \gamma)$ and $m_n^{(q)}(\phi, \gamma)$ be as defined in (25). Then note that with $u^+(\phi_0, \gamma) = \Omega(\gamma)^{-1}(u_1 - f_{10}u_2)$, the variance of $m_n^{(l)}(\phi_0, \gamma)$ is

$$\frac{1}{n} E\left[H'\Omega(\gamma)^{-1}\Omega(\gamma_0)H\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{\gamma_{(i),0}^2}{\gamma_{(i)}^2} h_i' h_i\right] \tag{28}$$

and the variance of $m_n^{(q)}(\phi_0, \gamma)$ is

$$\frac{1}{n} E\left[\text{tr}\left(\Omega_0\Omega^{-1}A\Omega_0\Omega^{-1}A\right)\right] = \frac{1}{n} E\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{\gamma_{(i),0}^2}{\gamma_{(i)}^2}\frac{\gamma_{(j),0}^2}{\gamma_{(j)}^2} a_{ij}^2\right]. \tag{29}$$

The assumption below ensures that for any admissible γ, Var $\left(m_n^{(l)}(\phi_0, \gamma)\right)$ and Var $\left(m_n^{(q)}(\phi_0, \gamma)\right)$ and their corresponding sample analogues converge, respectively, to finite positive definite matrices.

Assumption 8. *Let $h_i = [h_{i1}, ..., h_{iq}]$ denote the i-th row vector of H, and let a_{ij} denote the the (i, j)-th element of matrix A. The following holds :*

$$n^{-1}\sum_{i=1}^{n} E\left[\frac{\gamma_{(i),0}^2}{\gamma_{(i)}^2} h_i' h_i\right] \to V_\gamma^h, \quad n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left[\frac{\gamma_{(i),0}^2}{\gamma_{(i)}^2}\frac{\gamma_{(j),0}^2}{\gamma_{(j)}^2} a_{ij}^2\right] \to V_\gamma^a,$$

where the elements of V_γ^h and V_γ^a are finite , and

$$V_{n,\gamma}^h = n^{-1}\sum_{i=1}^{n}\frac{\gamma_{(i),0}^2}{\gamma_{(i)}^2} h_i' h_i \xrightarrow{p} V_\gamma^h, \quad V_{n,\gamma}^a = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\gamma_{(i),0}^2}{\gamma_{(i)}^2}\frac{\gamma_{(j),0}^2}{\gamma_{(j)}^2} a_{ij}^2 \xrightarrow{p} V_\gamma^a.$$

The matrix $V_\gamma = \text{diag}\left(V_\gamma^h, 2V_\gamma^a\right)$ is positive definite.

(i) Let $C(\theta, \gamma)$ be a $n \times n$ matrix of the form D, DP, DAD, $DADP$, or $PDADP$, where D is an $n \times n$ positive diagonal matrix with elements which are uniformly bounded and measurable w.r.t. $\mathcal{Z}_n$ and where P is $P = \text{diag}_c\left(pI_c^* + qJ_c^*\right)$ and where p and q are continuously differentiable functions of the parameters θ and γ as well as of variables generating $\mathcal{Z}_n$. Then

$$\lim_{n\to\infty}\sup_{\theta,\gamma}\left|n^{-1}\Upsilon_a' C(\theta, \gamma)\Upsilon_b - n^{-1}E\left[\Upsilon_a' C(\theta, \gamma)\Upsilon_b\right]\right| = 0 \ a.s.$$

where Υ_a, Υ_b are selected from the set $\{H, X, \mu^*, u_1, u_2\}$.

(ii) $\lim_{n\to\infty}\sup_{\theta,\gamma}\left\|E\left[n^{-1}\Upsilon_a' C(\theta, \gamma)\Upsilon_b\right] - \mathscr{U}_{a,b}(\theta, \gamma)\right\| = 0$ where $\mathscr{U}_{a,b}(\theta, \gamma)$ is bounded and continuously differentiable in θ and γ.

In Kuersteiner, Prucha, and Zeng (2023) we show uniform convergence under more primitive conditions in a closely related situation while directly assuming it here in Assumption 8 to save space. Without further assumptions on cross-sectional dependence, high level assumptions about the convergence of sample averages such as $n^{-1}\underline{X}'\underline{X}$ are required irrespective. The next theorem establishes consistency of the GMM estimator.

Theorem 4.1. *Let Assumptions 1- 8 be satisfied. Assume that $\bar{\gamma}_n \xrightarrow{p} \gamma$ for some sequence $\bar{\gamma}_n$ and some $\gamma \in \Gamma$. Then,*

(i) for $\hat{\phi}(\gamma) = \operatorname{argmin}_\theta Q_n(\phi, \gamma)$ with $Q_n(\phi, \gamma)$ defined in (15) it follows that $\hat{\phi}(\bar{\gamma}_n) \xrightarrow{p} \phi_0$.

(ii) for $\tilde{f}_1$ and $\tilde{\delta}$ defined in (16) *and $\tilde{\rho}$ defined in* (18) *it follows that $\tilde{f}_1 \xrightarrow{p} f_{0,1}$, $\tilde{\delta} \xrightarrow{p} \delta_0$ and $\tilde{\rho} \xrightarrow{p} \rho_0$.*

4.3 Asymptotic Normality of Parameter Estimators

In the following we develop an asymptotically justified inference theory for our proposed GMM estimator. For that purpose we first discuss consistent estimators for the variance parameters $\gamma = (\gamma_1^2, ..., \gamma_J^2)'$, which are treated as nuisance parameters in the objective function of our GMM estimator.

Assumption 9. *Let $\omega_j \equiv \lim_{n\to\infty} n^{-1} \sum_{c=1}^{C} n_c 1\{\tau_c = j\}$ and assume that ω_j exists and $\omega_j > 0$ for all $j = 1, ..., J$.*

In the following we denote with K_u a generic finite constant (which is taken, w.o.l.o.g., to be greater then one) and which is invariant over $t = 1, 2$, $i = 1, \ldots, n$, $n \geq 1$. The central limit theorem of Kuersteiner and Prucha (2020) requires that $E\left[|u_{it}u_{is}|^{1+\eta_u} | \mathcal{Z}_n\right] \leqslant K_u$ with $\eta_u > 0$ for $t, s = 1, 2$ and all i. This condition is implied by Assumption 2, as Lemma B.3 in the appendix demonstrates. Assumption 9 guarantees that there is enough data to estimate γ_j^2 for each j by requiring that the fraction ω_j of all students in class rooms of type j is asymptotically non-negligible for all types of class rooms.

The first step inefficient but consistent estimators $\tilde{f}_1$, $\tilde{\delta}$ and $\tilde{\rho}$ defined in (16) and (18) can be used to obtain consistent estimates of the variance parameters γ. Form the residuals

$$\tilde{\epsilon} = \epsilon^+(\tilde{\phi}) = (I_n + \tilde{\rho}M)^{-1} \left(y_1 - y_2\tilde{f}_1 - \underline{X}\tilde{\delta}\right) \tag{30}$$

and organize the residuals by class rooms c as $\tilde{\epsilon} = (\tilde{\epsilon}_1', ..., \tilde{\epsilon}_C')'$. Then construct the variance estimators

$$\hat{\gamma}_j^2 = (N_j - p_x - 1)^{-1} \sum_{c=1}^{C} \tilde{\epsilon}_c' \tilde{\epsilon}_c 1(\tau_c = j), \tag{31}$$

where $N_j = \sum_{c=1}^{C} n_c (\tau_c = j)$ is the number of students in class rooms of type j, and where p_x

denotes the number of columns in $\underline{X}$. Now set

$$\hat{\gamma} = \left(\hat{\gamma}_1^2, \ldots, \hat{\gamma}_J^2\right)'. \tag{32}$$

Efficient GMM estimators can now be formed by plugging $\hat{\gamma}$ into $u^+(\phi, \hat{\gamma})$ defined in (13). Recall that $\Omega(\gamma) = \operatorname{diag}_{c=1}^{C}(\gamma_{\tau_c}^2 I_c)$. The central limit theorem in Proposition 3 of Kuersteiner and Prucha (2020) can be applied to obtain distributional approximations for estimators based on u^+. When setting

$$\tilde{\gamma} = (1, 1, ..., 1) \tag{33}$$

it follows that $\Omega(\gamma) = I_n$. The feasible efficient GMM estimator for $\phi_0 = (\rho_0, f_{0,1}, \delta_0')'$ is defined as $\hat{\phi}\left(\hat{\gamma}\right) = \operatorname{argmin}_{\theta} Q_n\left(\phi, \hat{\gamma}\right)$ where $Q_n\left(\phi, \gamma\right)$ is given in (15) and the inefficient counterpart $\hat{\phi}\left(\tilde{\gamma}\right)$ is defined in an analogous way. The next theorem establishes the limiting distribution of $\hat{\phi}\left(\tilde{\gamma}\right)$ and $\hat{\phi}\left(\hat{\gamma}\right)$. The limiting distribution of $\hat{\phi}\left(\tilde{\gamma}\right)$ for $\tilde{\gamma}$ defined in (33) provides a distributional result for the first step estimators defined in (16) and (18).

Theorem 4.2. *Let Assumptions 1-9 hold. Let $\hat{\gamma}$ be as defined in* (32) *and $\tilde{\gamma}$ as defined in* (33)*. For $\gamma_n \to_p \gamma$ let $\operatorname{plim}_{n\to\infty} n^{-1/2}\partial m_n\left(\hat{\phi}\left(\gamma_n\right), \gamma_n\right)/\partial\phi = G(\gamma)$ and $\Xi = \operatorname{plim}_{n\to\infty} \Xi_n$, where Ξ_n is defined in* (14)*. Then $\hat{\gamma} \to_p \gamma_0$, and for $\gamma_n = \hat{\gamma}$ or $\gamma_n = \tilde{\gamma}$,*

$$\sqrt{n}\left(\hat{\phi}\left(\gamma_n\right) - \phi_0\right) \to_d N(0, \Psi_\gamma)$$

where the limiting variance covariance matrix Ψ_γ has the form

$$\Psi_\gamma = \left(G'\Xi^{-1}G\right)^{-1} G'\Xi^{-1}V_\gamma\Xi^{-1}G\left(G'\Xi^{-1}G\right)^{-1}.$$

with V_γ defined in Assumption 8 and G is the shorthand notation for $G(\gamma)$. When $\gamma_n = \hat{\gamma}$ then $V_\gamma = \Xi$ and $\Psi_\gamma = \left(G(\gamma_0)'\Xi^{-1}G(\gamma_0)\right)^{-1}$.

In the appendix in Section D we provide explicit formulas for the derivatives G. Together with the expressions for Ξ_n it is then easy to construct sample based estimators $\hat{G}$ and $\hat{\Xi}$. Letting $\hat{\Psi}_\gamma = \left(\hat{G}'\hat{\Xi}^{-1}\hat{G}\right)^{-1}$ it follows from a standard Slutsky argument that $\hat{\Psi}_\gamma^{-1/2}\sqrt{n}\left(\hat{\phi}\left(\hat{\gamma}\right) - \phi_0\right) \to_d N\left(0, I\right)$.

5 Conclusions

We identify peer effects using differential scores of closely related subject tests. Our analysis and identification strategy is based on a variance decomposition that distinguishes between test invariant components and unobserved idiosyncratic effects that are not predictable using concurrent test performance in related tests. The additional parametric assumption of linear education production functions and peer effects allows to identify a common endogenous peer effects parameter. Our method depends on the availability of closely related, yet separate test results but has the advantage of being robust to non-random group assignment. We prove identification of model parameters under conditions that are typical in the linear instrumental variables literature. These results depend on a careful exploration of the parametric structure of linear peer effects models. Our empirical results for Project STAR data and classrooms in Kindergarten to Third Grade show highly significant peer effects, but smaller in magnitude than those found in a number of studies that have looked at the same data.

References

Angrist, Joshua D. 2014. "The Perils of Peer Effects." *Labour Economics* 30:98–108.

Angrist, Joshua D. and Kevin Lang. 2004. "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program." *The American Economic Review* 94 (5):1613–1634.

Anselin, Luc. 2010. "Thirty Years of Spatial Econometrics." *Papers in Regional Science* 89 (1):3–25.

Arcidiacono, Peter, Gigi Foster, Natalie Goodpaster, and Josh Kinsler. 2012. "Estimating Spillovers Using Panel Data, with an Application to the Classroom." *Quantitative Economics* 3 (3):421–470.

Auerbach, Eric. 2022. "Identification and Estimation of a Partially Linear Regression Model Using Network Data." *Econometrica* 90 (1):347–365.

Bietenbeck, Jan. 2020. "The Long-Term Impacts of Low-Achieving Childhood Peers: Evidence from Project STAR." *Journal of the European Economic Association* 18:392–426.

Blume, Lawrence E., William A. Brock, Steven N. Durlauf, and Rajshri Jayaraman. 2015. "Linear Social Interactions Models." *Journal of Political Economy* 123 (2):444–496.

Booij, Adam S, Edwin Leuven, and Hessel Oosterbeek. 2017. "Ability Peer Effects in University: Evidence from a Randomized Experiment." *The Review of Economic Studies* 84 (2):547–578.

Boozer, Michael and Stephen E. Cacciola. 2001. "Inside the 'Black Box' of Project Star: Estimation of Peer Effects Using Experimental Data." SSRN Scholarly Paper, Social Science Research Network, Rochester, NY.

Boyd-Zaharias, J, J Finn, R Fish, and S Gerber. 2007. "Project STAR and Beyond: Database Users Guide." Tech. rep., Tech. rep.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. 2009. "Identification of Peer Effects through Social Networks." *Journal of Econometrics* 150 (1):41–55.

Burke, Mary A and Tim R Sass. 2013. "Classroom Peer Effects and Student Achievement." *Journal of Labor Economics* 31 (1):51–82.

Calvó-Armengol, Antoni, Eleonora Patacchini, and Yves Zenou. 2009. "Peer Effects and Social Networks in Education." *The Review of Economic Studies* 76 (4):1239–1267.

Carrell, Scott E., Mark Hoekstra, and Elira Kuka. 2018. "The Long-Run Effects of Disruptive Peers." *American Economic Review* 108 (11):3377–3415.

Carrell, Scott E., Bruce I. Sacerdote, and James E. West. 2013. "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation." *Econometrica* 81 (3):855–882.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *The Quarterly Journal of Economics* 126 (4):1593–1660.

Cipollone, Piero and Alfonso Rosolia. 2007. "Social Interactions in High School: Lessons from an Earthquake." *American Economic Review* 97 (3):948–965.

Cliff, Andrew David and J Keith Ord. 1973. *Spatial Autocorrelation*, vol. 5. Pion London.

———. 1981. *Spatial Processes: Models & Applications*, vol. 44. Pion London.

Cohen-Cole, Ethan, Xiaodong Liu, and Yves Zenou. 2018. "Multivariate Choices and Identification of Social Interactions." *Journal of Applied Econometrics* 33 (2):165–178.

Cornelissen, Thomas, Christian Dustmann, and Uta Schönberg. 2017. "Peer Effects in the Workplace." *American Economic Review* 107 (2):425–456.

Dee, Thomas S. 2004. "Teachers, Race, and Student Achievement in a Randomized Experiment." *Review of Economics and Statistics* 86 (1):195–210.

Drukker, David M., Peter H. Egger, and Ingmar R. Prucha. 2022. "SIMULTANEOUS EQUATIONS MODELS WITH HIGHER-ORDER SPATIAL OR SOCIAL NETWORK INTERACTIONS." *Econometric Theory* :1–48.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *The American Economic Review* 101 (5):1739–1774.

Duflo, Esther and Emmanuel Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment*." *The Quarterly Journal of Economics* 118 (3):815–842.

Feld, Jan and Ulf Zölitz. 2016. "Understanding Peer Effects: On the Nature, Estimation, and Channels of Peer Effects." *Journal of Labor Economics* 35 (2):387–428.

Garlick, Robert. 2018. "Academic Peer Effects with Different Group Assignment Policies: Residential Tracking versus Random Assignment." *American Economic Journal: Applied Economics* 10 (3):345–369.

Goldsmith-Pinkham, Paul and Guido W. Imbens. 2013. "Social Networks and the Identification of Peer Effects." *Journal of Business & Economic Statistics* 31 (3):253–264.

Graham, Bryan S. 2008. "Identifying Social Interactions Through Conditional Variance Restrictions." *Econometrica* 76 (3):643–660.

Griffith, Alan. 2022. "Random Assignment with Non-Random Peers: A Structural Approach to Counterfactual Treatment Assessment." *Review of Ecnomics and Statistics* (forthcoming).

Hanushek, Eric A. 1999. "Some Findings From an Independent Investigation of the Tennessee STAR Experiment and From Other Investigations of Class Size Effects." *Educational Evaluation and Policy Analysis* 21 (2):143–163.

———. 2003. "The Failure of Input-based Schooling Policies*." *The Economic Journal* 113 (485):F64–F98.

Hanushek, Eric A., John F. Kain, Jacob M. Markman, and Steven G. Rivkin. 2003. "Does Peer Ability Affect Student Achievement?" *Journal of Applied Econometrics* 18 (5):527–544.

Hoxby, Caroline. 2000. "Peer Effects in the Classroom: Learning from Gender and Race Variation." Working Paper, National Bureau of Economic Research.

Hsieh, Chih-Sheng and Lung Fei Lee. 2016. "A Social Interactions Model with Endogenous Friendship Formation and Selectivity." *Journal of Applied Econometrics* 31:301–319.

Johnsson, Ida and Hyungsik Roger Moon. 2019. "Estimation of Peer Effects in Endogenous Social Networks: Control Function Approach." *The Review of Economics and Statistics* 103 (2):328–345.

Kelejian, Harry H. and Ingmar R. Prucha. 2002. "2SLS and OLS in a Spatial Autoregressive Model with Equal Spatial Weights." *Regional Science and Urban Economics* 32 (6):691–707.

———. 2004. "Estimation of Simultaneous Systems of Spatially Interrelated Cross Sectional Equations." *Journal of Econometrics* 118 (1–2):27–50.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics* 114 (2):497–532.

Kuersteiner, Guido M. and Ingmar R. Prucha. 2020. "Dynamic Spatial Panel Models: Networks, Common Shocks, and Sequential Exogeneity." *Econometrica* 88 (5):2109–2146.

Kuersteiner, Guido M., Ingmar R. Prucha, and Ying Zeng. 2023. "Efficient Peer Effects Estimators with Random Group Effects." *forthcoming Journal of Econometrics* .

Lazear, Edward P. 2001. "Educational Production." *The Quarterly Journal of Economics* 116 (3):777–803.

Lee, Lung-Fei. 2007. "Identification and Estimation of Econometric Models with Group Interactions, Contextual Factors and Fixed Effects." *Journal of Econometrics* 140 (2):333–374.

Lewbel, Arthur, Xi Qu, and Xun Tang. 2023. "Social Networks with Unobserved Links." *Journal of Political Economy* 131:898–946.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *The Review of Economic Studies* 60 (3):531–542.

Mas, Alexandre and Enrico Moretti. 2009. "Peers at Work." *The American Economic Review* 99 (1):112–145.

Miraldo, Marisa, Carol Propper, and Christiern Rose. 2021. "Identification of Peer Effects Using Panel Data." .

Mosteller, Frederick. 1995. "The Tennessee Study of Class Size in the Early School Grades." *The Future of Children* 5 (2):113–127.

Mueller, Steffen. 2013. "Teacher Experience and the Class Size Effect — Experimental Evidence." *Journal of Public Economics* 98:44–52.

Newey, Whitney K. and Daniel McFadden. 1994. "Chapter 36 Large sample estimation and hypothesis testing." In *Handbook of Econometrics*, vol. 4. Elsevier, 2111–2245. URL `https://linkinghub.elsevier.com/retrieve/pii/S1573441205800054`.

Pereda-Fernández, Santiago. 2017. "Social Spillovers in the Classroom: Identification, Estimation and Policy Analysis." *Economica* 84 (336):712–747.

Pötscher, Benedikt M. and Ingmar Prucha. 1997. *Dynamic Nonlinear Econometric Models: Asymptotic Theory*. Springer Science & Business Media.

Qu, Xi and Lung-fei Lee. 2015. "Estimating a Spatial Autoregressive Model with an Endogenous Spatial Weight Matrix." *Journal of Econometrics* 184 (2):209–232.

Rose, Christiern D. 2017. "Identification of Peer Effects through Social Networks Using Variance Restrictions." *The Econometrics Journal* 20 (3):S47–S60.

Rothenberg, Thomas J. 1971. "Identification in Parametric Models." *Econometrica* 39 (3):577–591. URL `https://www.jstor.org/stable/1913267`. Publisher: [Wiley, The Econometric Society].

Sacerdote, Bruce. 2001. "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *The Quarterly Journal of Economics* 116 (2):681–704.

———. 2011. "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?" In *Handbook of the Economics of Education*, vol. Volume 3, edited by Stephen Machin and Ludger Woessmann Eric A. Hanushek. Elsevier, 249–277.

Sojourner, Aaron. 2013. "Identification of Peer Effects with Missing Peer Data: Evidence from Project STAR." *The Economic Journal* 123 (569):574–605.

Todd, Petra E. and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement*." *The Economic Journal* 113 (485):F3–F33.

Whitmore, Diane. 2005. "Resource and Peer Impacts on Girls' Academic Achievement: Evidence from a Randomized Experiment." *The American Economic Review* 95 (2):199–203.

Word, Elizabeth, John Johnston, Helen P. Bain, B. DeWayne Fulton, Jayne B. Zaharias, Charles M. Achilles, Martha N. Lintz, John Folger, and Carolyn Breda. 1990. "The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report(1985-1990)." *Tennessee Board of Education* .

Wu, Jia, Junsen Zhang, and Chunchao Wang. 2023. "Student Performance, Peer Effects, and Friend Networks: Evidence from a Randomized Peer Intervention." 15:510–542.

Zimmerman, David J. 2003. "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment." *Review of Economics and Statistics* 85 (1):9–23.

A Tables

Table 1: Summary Statistics

	Grade K		Grade 1		Grade 2		Grade 3	
	mean	sd	mean	sd	mean	sd	mean	sd
				indiviudal characteristics				
free lunch	0.560	0.496	0.588	0.492	0.592	0.491	0.578	0.494
black	0.326	0.469	0.327	0.469	0.347	0.476	0.333	0.471
girl	0.486	0.500	0.480	0.500	0.484	0.500	0.480	0.500
age	6.011	0.349	7.158	0.492	8.242	0.557	9.285	0.586
				teacher characteristics				
black teacher	0.165	0.371	0.174	0.379	0.204	0.403	0.209	0.406
master degree	0.347	0.476	0.346	0.476	0.373	0.484	0.442	0.497
years of experience	9.258	5.809	11.633	8.937	13.145	8.655	13.933	8.615
				SAT test scores				
math	485.377	47.698	530.528	43.109	580.613	44.574	617.970	39.841
read	436.725	31.706	520.787	55.192	583.935	46.043	615.422	38.563
list	537.475	33.140	567.487	33.674	595.476	34.908	624.119	32.265
word	434.179	36.759	513.436	53.316	582.986	50.658	610.136	45.041
				Sample Size				
number of students	6325		6829		6840		6801	
number of classes	325		339		340		336	
# small	127		124		133		140	
# regular	99		115		100		89	
# regular/aide	99		100		107		107	
number of schools	79		76		75		75	

1. Mean and standard deviations of student characteristics, teacher charcteristics, and SAT scores, as well as number of students and classes in each grade.

Table 2: Rank Correlation between Scores

		Rank Correlation between Tests		Pseudo R squared of 2SLS	
y_1	y_2	y_1, y_2	$Q_X y_1, Q_X y_2$	$y_1, \hat{y}_1$	$y_2, \hat{y}_2$
			Grade K		
list	math	0.657	0.584	0.638	0.728
list	read	0.621	0.554	0.691	0.686
list	word	0.558	0.477	0.610	0.637
read	math	0.741	0.672	0.697	0.763
word	math	0.656	0.579	0.643	0.727
word	read	0.908	0.881	0.915	0.909
			Grade 1		
list	math	0.721	0.654	0.716	0.766
list	read	0.624	0.534	0.601	0.704
list	word	0.570	0.483	0.562	0.645
read	math	0.737	0.669	0.782	0.753
word	math	0.686	0.619	0.726	0.715
word	read	0.932	0.905	0.936	0.940
			Grade 2		
list	math	0.678	0.611	0.705	0.713
list	read	0.667	0.601	0.699	0.709
list	word	0.597	0.533	0.639	0.639
read	math	0.724	0.673	0.756	0.758
word	math	0.625	0.564	0.659	0.670
word	read	0.884	0.848	0.892	0.894
			Grade 3		
list	math	0.649	0.610	0.673	0.689
list	read	0.652	0.623	0.683	0.690
list	word	0.524	0.478	0.552	0.581
read	math	0.741	0.706	0.759	0.769
word	math	0.683	0.643	0.712	0.708
word	read	0.890	0.869	0.901	0.895

1. The left panel is Spearman's rank correlation between y_1 y_2 and between $Q_X y_1$ and $Q_X y_2$, where $Q_X = I - X(X'X)^{-1}X'$, y_1 and y_2 are raw SAT scores specified in Columns 1 and 2, X are the complete set of control variables, including school fixed effects, class type fixed effects, student characteristics, teacher characteristics and peer characteristics.
2. The right panel is pseudo R^2, i.e., Spearman's rank correlation between y_1 and $\hat{y}_1$ in Column 3, and between y_2 and $\hat{y}_2$ in Column 4, where $\hat{y}_1$ is the predicted value of y_1 from 2SLS for $y_1 = f_1 y_2 + X'\beta + \epsilon$ with y_2 instrumented by the constant term, X is the complete set of controls as described in note 1, $\hat{y}_2$ is defined in a similar manner by reversing the order of y_1, y_2.

Table 3: Explanatory Power of Lagged Scores

lagged score	Estiamtes for lagged scores Grade 1	Grade 2	Grade 3	R^2 w/wo lagged scores Grade 1	Grade 2	Grade 3
	Dependent Variable: Listening Score					
N.A.				0.256	0.297	0.236
list	**0.620(0.019)**	**0.707(0.014)**	**0.687(0.013)**	**0.515**	**0.617**	**0.599**
math	0.362(0.012)	0.465(0.013)	0.394(0.011)	0.435	0.488	0.419
read	0.476(0.020)	0.303(0.009)	0.386(0.011)	0.410	0.448	0.424
word	0.355(0.016)	0.272(0.010)	0.296(0.010)	0.381	0.409	0.378
	Dependent Variable: Mathematics Score					
N.A.				0.284	0.272	0.258
list	0.616(0.021)	0.679(0.019)	0.606(0.017)	0.454	0.464	0.444
math	**0.549(0.015)**	**0.780(0.016)**	**0.697(0.013)**	**0.537**	**0.614**	**0.630**
read	0.655(0.026)	0.455(0.011)	0.554(0.013)	0.464	0.496	0.513
word	0.505(0.019)	0.421(0.014)	0.396(0.012)	0.442	0.455	0.428
	Dependent Variable: Reading Score					
N.A.				0.290	0.294	0.246
list	0.592(0.026)	0.662(0.020)	0.628(0.016)	0.391	0.451	0.458
math	0.592(0.019)	0.669(0.018)	0.538(0.013)	0.468	0.521	0.485
read	**1.026(0.032)**	**0.622(0.011)**	**0.671(0.010)**	**0.544**	**0.648**	**0.644**
word	0.809(0.024)	0.582(0.013)	0.512(0.010)	0.522	0.588	0.544
	Dependent Variable: Word Study Skills Score					
N.A.				0.244	0.250	0.229
list	0.509(0.028)	0.645(0.022)	0.579(0.020)	0.333	0.377	0.363
math	0.511(0.020)	0.645(0.020)	0.579(0.016)	0.399	0.426	0.441
read	0.860(0.033)	0.668(0.012)	0.737(0.013)	0.453	0.588	0.594
word	**0.686(0.025)**	**0.631(0.014)**	**0.547(0.012)**	**0.435**	**0.539**	**0.488**

1. Each panel represents regressions of a specific score types against a comprehensive set of control variables (including school fixed effects, class type fixed effects, student, teacher, and peer characteristics in the main specification and a constant term), along with possible own and peer average lagged scores.
2. The lagged score type is indicated in column 1 and highlighted in bold if it matches the dependent variable type. N.A. means no lagged scores are controlled for.
3. The left panel displays coefficients and standard deviations (in parentheses) of lagged scores for grade 1 to 3.The right panel reports the R^2 of the regression, where the first row in each panel representing the R^2 when no lagged scores are included.
4. Coefficients of peer average lagged scores, typically small and statistically insignificant, are omitted for conciseness.

Table 4: Pseudo R^2: $1 - var(u_1 - f_1 u_2)/var(y_1 - f_1 y_2)$

		Small Classes				Regualr Classes(w/wo Aide)			
y1	y2	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
					Grade K				
list	math	0.179	0.180	0.206	0.216	0.177	0.179	0.201	0.208
list	read	0.143	0.144	0.153	0.150	0.145	0.146	0.150	0.147
list	word	0.124	0.125	0.132	0.129	0.132	0.133	0.136	0.134
math	read	0.148	0.148	0.177	0.219	0.126	0.127	0.155	0.192
math	word	0.138	0.139	0.164	0.200	0.128	0.129	0.154	0.184
read	word	0.104	0.104	0.107	0.113	0.140	0.140	0.142	0.146
					Grade 1				
list	math	0.138	0.140	0.143	0.144	0.113	0.115	0.123	0.124
list	read	0.131	0.133	0.154	0.163	0.113	0.116	0.148	0.161
list	word	0.105	0.106	0.122	0.132	0.068	0.071	0.094	0.108
math	read	0.105	0.105	0.130	0.137	0.070	0.071	0.119	0.132
math	word	0.092	0.092	0.114	0.118	0.058	0.059	0.101	0.109
read	word	0.088	0.088	0.092	0.090	0.093	0.093	0.098	0.093
					Grade 2				
list	math	0.116	0.117	0.123	0.119	0.104	0.106	0.118	0.109
list	read	0.087	0.087	0.144	0.141	0.083	0.084	0.125	0.125
list	word	0.084	0.084	0.131	0.132	0.070	0.072	0.107	0.109
math	read	0.092	0.092	0.134	0.137	0.086	0.087	0.110	0.109
math	word	0.089	0.089	0.124	0.126	0.081	0.081	0.106	0.107
read	word	0.064	0.064	0.077	0.082	0.065	0.065	0.074	0.078
					Grade 3				
list	math	0.135	0.136	0.147	0.144	0.085	0.087	0.092	0.089
list	read	0.073	0.074	0.133	0.133	0.024	0.027	0.077	0.078
list	word	0.090	0.090	0.132	0.131	0.030	0.033	0.065	0.066
math	read	0.083	0.083	0.107	0.106	0.050	0.050	0.085	0.086
math	word	0.085	0.085	0.107	0.106	0.039	0.040	0.071	0.070
read	word	0.093	0.093	0.092	0.096	0.033	0.034	0.035	0.038
school FE		Y	Y	Y	Y	Y	Y	Y	Y
class type FE			Y	Y	Y		Y	Y	Y
stu&tch char				Y	Y			Y	Y
peer char					Y				Y

1. The table reports one minus the ratio of the variance of $u_1 - f_1 u_2$ to that of $y_1 - f_1 y_2$ for small classes (in the left panel) and regular (with or without aide) classes (in the right panel)All models allow for heteroscedasticity across small and regular classes.

2. The scores are raw SAT scores in listening (list), mathematics (math), reading (read), word study skills (word).

3. As indicated at the bottom of the table, the models may control for (1) school fixed effects (excluding one); (2) class type fixed effects (excluding small); (3) a student's characteristics (free lunch, black, girl, age) and teacher characteristics (black teacher, master, years of experience); (4) peer averages of student characteristics.

Table 5: Standard Deviations of $Q_x y$ and $Q_x(y_1 - f_1 y_2)$

	y_1	y_2	Small Classes (1)	Small Classes (2)	Small Classes (3)	Small Classes (4)	Regular Classes (1)	Regular Classes (2)	Regular Classes (3)	Regular Classes (4)
Grade K										
$Q_x y$	list		75.810	75.122	37.769	28.508	78.091	76.238	38.954	29.681
$Q_x y$	math		75.456	74.762	44.444	39.421	80.333	78.669	48.468	43.319
$Q_x y$	read		62.748	62.169	33.699	26.827	66.321	64.827	35.534	29.143
$Q_x y$	word		65.196	64.629	38.017	31.883	68.723	67.294	39.693	33.916
$Q_x(y_1 - f_1 y_2)$	list	math	35.462	35.555	38.504	41.091	38.064	38.131	41.515	44.471
$Q_x(y_1 - f_1 y_2)$	list	read	30.802	30.801	30.488	27.544	31.701	31.697	31.347	28.060
$Q_x(y_1 - f_1 y_2)$	list	word	36.916	36.919	36.311	33.052	37.006	37.015	36.336	32.868
$Q_x(y_1 - f_1 y_2)$	math	read	30.874	30.867	30.338	30.583	31.934	31.927	31.556	32.405
$Q_x(y_1 - f_1 y_2)$	math	word	35.724	35.696	34.235	32.883	35.731	35.707	34.486	33.819
$Q_x(y_1 - f_1 y_2)$	read	word	14.055	14.059	13.883	14.896	14.625	14.632	14.422	15.448
Grade 1										
$Q_x y$	list		78.430	77.682	44.281	29.291	82.153	80.072	43.729	32.263
$Q_x y$	math		77.259	76.454	47.817	36.254	81.842	79.872	48.874	39.564
$Q_x y$	read		83.720	82.984	57.172	46.261	88.372	86.584	58.298	49.448
$Q_x y$	word		81.314	80.624	57.574	46.620	85.880	84.192	57.174	47.750
$Q_x(y_1 - f_1 y_2)$	list	math	29.766	29.822	29.377	29.449	30.662	30.730	30.227	30.285
$Q_x(y_1 - f_1 y_2)$	list	read	42.790	42.812	38.711	40.458	43.672	43.660	39.169	40.844
$Q_x(y_1 - f_1 y_2)$	list	word	44.810	44.856	40.579	42.694	45.258	45.272	40.989	42.897
$Q_x(y_1 - f_1 y_2)$	math	read	35.658	35.595	32.807	34.226	37.625	37.547	34.073	35.289
$Q_x(y_1 - f_1 y_2)$	math	word	38.230	38.195	35.500	36.668	39.197	39.143	36.076	36.874
$Q_x(y_1 - f_1 y_2)$	read	word	21.646	21.649	21.648	21.786	22.530	22.533	22.493	22.616
Grade 2										
$Q_x y$	list		92.075	90.731	48.088	29.859	85.236	82.708	45.789	32.371
$Q_x y$	math		90.878	89.540	53.323	38.275	85.553	83.117	52.234	41.389
$Q_x y$	read		93.586	92.295	57.595	39.409	87.831	85.406	54.519	41.740
$Q_x y$	word		96.864	95.555	62.274	44.613	90.149	87.782	59.070	46.967
$Q_x(y_1 - f_1 y_2)$	list	math	32.791	32.843	31.327	28.848	34.666	34.691	32.803	30.147
$Q_x(y_1 - f_1 y_2)$	list	read	34.210	34.242	29.824	28.460	34.229	34.216	29.939	28.515
$Q_x(y_1 - f_1 y_2)$	list	word	40.464	40.509	34.546	31.816	39.898	39.907	34.124	31.404
$Q_x(y_1 - f_1 y_2)$	math	read	31.591	31.571	29.607	30.029	32.541	32.523	31.102	31.550
$Q_x(y_1 - f_1 y_2)$	math	word	39.107	39.083	35.878	35.793	39.412	39.388	36.700	36.676
$Q_x(y_1 - f_1 y_2)$	read	word	23.891	23.894	23.273	22.593	24.312	24.318	23.726	23.043
Grade 3										
$Q_x y$	list		100.934	98.926	46.812	28.881	81.116	78.549	45.571	31.663
$Q_x y$	math		99.438	97.593	51.961	35.145	82.400	79.955	50.671	37.816
$Q_x y$	read		102.223	100.223	52.813	33.948	85.025	82.325	51.772	38.114
$Q_x y$	word		102.419	100.495	56.860	39.320	86.959	84.387	56.262	43.608
$Q_x(y_1 - f_1 y_2)$	list	math	29.971	29.941	28.331	27.089	30.284	30.259	28.579	27.304
$Q_x(y_1 - f_1 y_2)$	list	read	29.166	29.145	26.466	25.715	29.828	29.770	26.962	26.078
$Q_x(y_1 - f_1 y_2)$	list	word	38.217	38.212	34.066	33.314	39.272	39.206	34.788	33.930
$Q_x(y_1 - f_1 y_2)$	math	read	27.010	27.013	26.326	26.477	26.807	26.801	25.922	26.045
$Q_x(y_1 - f_1 y_2)$	math	word	32.785	32.796	31.350	31.918	32.844	32.832	31.064	31.683
$Q_x(y_1 - f_1 y_2)$	read	word	20.631	20.634	20.090	20.639	21.111	21.106	20.485	21.043

1. Standard deviations of $Q_x y$ and $Q_x(y_1 - f_1 y_2)$ for small (on the left) and regular classes (on the right). For $Q_x y$, the test score y is specified in y_1. For $Q_x(y_1 - f_1 y_2)$, y_1 and y_2 are specifed in columns 2 and 3. The control variables for models (1)-(4) are the same as those for the main models in Tables 6 and 7. The estiamtes of f_1 are from our efficient GMM estimations for each specifications and presetned in Tables 6.

Table 6: Estiamtes of f_1

y_1	y_2	(1)	(2)	(3)	(4)
			Grade K		
list	math	1.093(0.011)	1.096(0.010)	1.201(0.027)	1.283(0.243)
list	read	1.231(0.008)	1.231(0.008)	1.217(0.021)	1.011(0.109)
list	word	1.224(0.007)	1.225(0.006)	1.201(0.023)	1.048(0.128)
math	read	1.126(0.005)	1.124(0.004)	1.014(0.018)	0.771(0.140)
math	word	1.120(0.008)	1.118(0.008)	0.999(0.021)	0.793(0.152)
read	word	0.995(0.004)	0.995(0.005)	0.984(0.009)	1.042(0.055)
			Grade 1		
list	math	1.065(0.009)	1.068(0.008)	1.054(0.012)	1.057(0.070)
list	read	1.060(0.008)	1.061(0.008)	0.973(0.014)	1.023(0.077)
list	word	1.071(0.009)	1.073(0.009)	0.977(0.015)	1.038(0.087)
math	read	0.995(0.005)	0.993(0.006)	0.916(0.012)	0.969(0.072)
math	word	1.007(0.004)	1.006(0.005)	0.929(0.012)	0.973(0.084)
read	word	1.008(0.005)	1.008(0.006)	1.014(0.009)	1.026(0.045)
			Grade 2		
list	math	1.057(0.005)	1.059(0.005)	1.007(0.012)	0.902(0.055)
list	read	1.030(0.006)	1.031(0.006)	0.903(0.009)	0.842(0.050)
list	word	1.023(0.006)	1.024(0.007)	0.873(0.011)	0.779(0.052)
math	read	0.975(0.006)	0.974(0.006)	0.893(0.010)	0.929(0.048)
math	word	0.967(0.006)	0.967(0.007)	0.863(0.010)	0.862(0.055)
read	word	0.992(0.004)	0.993(0.004)	0.967(0.007)	0.929(0.030)
			Grade 3		
list	math	1.028(0.002)	1.027(0.002)	0.961(0.008)	0.898(0.055)
list	read	1.014(0.003)	1.013(0.003)	0.915(0.008)	0.869(0.045)
list	word	1.027(0.004)	1.026(0.004)	0.903(0.010)	0.873(0.053)
math	read	0.986(0.003)	0.986(0.003)	0.951(0.008)	0.965(0.047)
math	word	0.999(0.004)	0.999(0.004)	0.940(0.009)	0.970(0.051)
read	word	1.013(0.002)	1.013(0.002)	0.988(0.005)	1.016(0.032)
school FE		Y	Y	Y	Y
class type FE			Y	Y	Y
stu&tch char				Y	Y
peer char					Y

1. Estimates and standard errors (in the parenthesis) for ρ. Standardard errors are clustered at the classroom level. All models allow for heteroscedasticity across small and regular classes. Estimates are adjusted for missing observations.
2. The scores are raw SAT scores in listening (list), mathematics (math), reading (read), word study skills(word).
3. As indicated at the bottom of the table, the models may control for (1) school fixed effects (excluding one); (2) class type fixed effects (excluding small); (3) a student's characteristics (free lunch, black, girl, age) and teacher characteristics (black teacher, master, years of experience); (4) peer averages of student characteristics.

Table 7: Estiamtes of ρ

y_1	y_2	(1)	(2)	(3)	(4)
		Grade K			
list	math	0.721(0.079)	0.691(0.073)	0.694(0.075)	0.688(0.078)
list	read	0.434(0.079)	0.415(0.076)	0.407(0.078)	0.377(0.083)
list	word	0.398(0.066)	0.378(0.065)	0.373(0.065)	0.359(0.068)
math	read	0.492(0.065)	0.485(0.065)	0.519(0.066)	0.575(0.076)
math	word	0.517(0.067)	0.513(0.067)	0.538(0.070)	0.578(0.081)
read	word	0.452(0.078)	0.449(0.078)	0.443(0.080)	0.424(0.075)
		Grade 1			
list	math	0.627(0.069)	0.587(0.073)	0.579(0.073)	0.578(0.074)
list	read	0.322(0.051)	0.277(0.049)	0.256(0.048)	0.218(0.046)
list	word	0.385(0.054)	0.340(0.052)	0.308(0.048)	0.257(0.047)
math	read	0.441(0.063)	0.434(0.064)	0.434(0.061)	0.360(0.061)
math	word	0.487(0.066)	0.480(0.065)	0.469(0.062)	0.388(0.064)
read	word	0.178(0.054)	0.178(0.054)	0.190(0.054)	0.193(0.053)
		Grade 2			
list	math	0.637(0.073)	0.621(0.072)	0.554(0.071)	0.525(0.068)
list	read	0.568(0.073)	0.553(0.071)	0.571(0.074)	0.561(0.074)
list	word	0.537(0.070)	0.521(0.068)	0.512(0.070)	0.497(0.069)
math	read	0.481(0.071)	0.479(0.071)	0.475(0.066)	0.444(0.066)
math	word	0.493(0.071)	0.492(0.071)	0.478(0.066)	0.459(0.065)
read	word	0.399(0.065)	0.398(0.064)	0.363(0.063)	0.348(0.064)
		Grade 3			
list	math	0.747(0.088)	0.728(0.087)	0.714(0.086)	0.703(0.090)
list	read	0.606(0.091)	0.572(0.088)	0.613(0.096)	0.623(0.101)
list	word	0.512(0.095)	0.482(0.091)	0.518(0.095)	0.522(0.097)
math	read	0.462(0.078)	0.460(0.078)	0.452(0.082)	0.440(0.086)
math	word	0.371(0.074)	0.366(0.072)	0.370(0.074)	0.357(0.078)
read	word	0.391(0.078)	0.387(0.077)	0.377(0.077)	0.379(0.076)
school FE		Y	Y	Y	Y
class type FE			Y	Y	Y
stu&tch char				Y	Y
peer char					Y

1. Estimates and standard errors (in the parenthesis) for ρ. Standardard errors are clustered at the classroom level. All models allow for heteroscedasticity across small and regular classes. Estimates are adjusted for missing observations.
2. The scores are raw SAT scores in listening (list), mathematics (math), reading (read), word study skills(word).
3. As indicated at the bottom of the table, the models may control for (1) school fixed effects (excluding one); (2) class type fixed effects (excluding small); (3) a student's characteristics (free lunch, black, girl, age) and teacher characteristics (black teacher, master, years of experience); (4) peer averages of student characteristics.

Table 8: Estiamtes of Exogenous Peer Effects

y_1	y_2	age	black(pct.)	free lunch(pct.)	girl(pct.)
			Grade K		
list	math	-4.290(15.730)	-0.196(0.244)	0.117(0.099)	-0.177(0.102)*
list	read	13.784(6.951)**	0.093(0.103)	0.024(0.050)	-0.031(0.050)
list	word	10.097(8.239)	0.066(0.134)	0.014(0.058)	0.020(0.058)
math	read	15.222(9.027)*	0.238(0.159)	-0.076(0.065)	0.111(0.063)*
math	word	12.705(9.930)	0.200(0.156)	-0.082(0.066)	0.159(0.064)**
read	word	-4.298(3.626)	-0.029(0.062)	-0.005(0.024)	0.070(0.027)**
			Grade 1		
list	math	-0.268(5.134)	-0.016(0.095)	-0.049(0.068)	0.013(0.079)
list	read	-1.922(6.076)	-0.125(0.124)	0.129(0.084)	-0.354(0.095)***
list	word	-2.212(6.848)	-0.186(0.147)	0.160(0.095)*	-0.454(0.108)***
math	read	-2.392(5.540)	-0.099(0.103)	0.179(0.076)**	-0.330(0.096)***
math	word	-1.156(6.399)	-0.161(0.118)	0.196(0.086)**	-0.416(0.107)***
read	word	-0.719(3.423)	-0.033(0.065)	0.038(0.043)	-0.037(0.050)
			Grade 2		
list	math	7.639(4.183)*	0.166(0.123)	-0.045(0.071)	0.120(0.089)
list	read	4.909(4.197)	0.109(0.104)	0.127(0.072)*	0.063(0.082)
list	word	7.808(4.551)*	-0.001(0.102)	0.150(0.079)*	0.149(0.086)*
math	read	-2.902(3.933)	-0.062(0.107)	0.187(0.068)***	-0.064(0.084)
math	word	-0.032(4.720)	-0.183(0.130)	0.213(0.082)***	0.025(0.096)
read	word	2.977(2.691)	-0.134(0.088)	0.025(0.050)	0.105(0.054)*
			Grade 3		
list	math	5.041(4.171)	0.187(0.094)**	-0.111(0.055)**	-0.059(0.066)
list	read	4.026(3.618)	0.030(0.084)	-0.079(0.048)*	-0.055(0.059)
list	word	2.712(4.286)	0.061(0.097)	-0.045(0.058)	-0.063(0.073)
math	read	-1.133(3.779)	-0.162(0.087)*	0.032(0.047)	0.003(0.054)
math	word	-2.395(4.101)	-0.129(0.104)	0.073(0.057)	-0.009(0.058)
read	word	-2.190(2.593)	0.030(0.062)	0.037(0.040)	-0.025(0.045)

1. Estimates and standard errors (in the parenthesis) for exogenoous peer effects of peer's age, race, free lunch status and gender. Coeffiicnets for dummies black, free lunch and girl are divided by 100 so they can be interpreted as the impact of one percentage point changes in these variables. Note that these four variables are all test-invariant personal characteristics. The reported coefficients are $\rho(\beta^p_{v1} - f_1\beta^p_{v2})$.
2. Standardard errors are clustered at the classroom level. All models allow for heteroscedasticity across small and regular classes. Estimates are adjusted for missing observations.
3. The scores are raw SAT scores in listening (list), mathematics (math), reading (read), word study skills(word).
4. All specifications control for (1) school fixed effects (excluding one); (2) class type fixed effects (excluding small); (3) a student's characteristics (free lunch, black, girl, age) and teacher characteristics (black teacher, master, years of experience); (4) peer averages of student characteristics.

B Auxiliary Lemmas

Before proving the main results, we establish a number of preliminary lemmas. Note that many matrices in this paper, e.g., M, $\Omega(\gamma)$, Σ_t defined in Assumption 3 can be written in the form of $\text{diag}_{c=1}^{C}\left(p_c I_c^* + q_c J_c^*\right)$,where $I_c^* = I_c - \mathbf{1}_c\mathbf{1}_c'/n_c$, $J_c^* = \mathbf{1}_c\mathbf{1}_c'/n_c$ are the residual projection matrix and the projection matrix onto $\mathbf{1}_c$ respectively. To see this note that

$$M = \text{diag}_{c=1}^{C}\left(-\frac{1}{n_c - 1}I_c^* + J_c^*\right), \tag{34}$$

$$I + \rho M = \text{diag}_{c=1}^{C}\left(\frac{n_c - 1 - \rho}{n_c - 1}I_c^* + (1+\rho)\,J_c^*\right), \tag{35}$$

$$\Omega(\gamma) = \text{diag}_{c=1}^{C}\left(\gamma_{\tau_c}^2 I_c\right) = \text{diag}_{c=1}^{C}\left(\gamma_{\tau_c}^2 I_c^* + \gamma_{\tau_c}^2 J_c^*\right), \tag{36}$$

$$\Sigma_t = \sigma_{0,t}^2\,\text{diag}_{c=1}^{C}\left(\rho_{0,\tau_c}^2 I_c\right) = \text{diag}_{c=1}^{C}\left(\rho_{0,\tau_c}^2\sigma_{0,t}^2 I_c^* + \rho_{0,\tau_c}^2\sigma_{0,t}^2 J_c^*\right),\ t = 1, 2, \tag{37}$$

We call the row and column sums of an $n \times n$ matrix $S(\phi, \gamma)$ with elements $s_{ij}(\phi, \gamma)$ uniformly bounded in absolute value if $\sup_{i,\phi,\gamma}\sum_{j=1}^{n}|s_{ij}(\phi,\gamma)| \leqslant C_s < \infty$ and $\sup_{j,\phi,\gamma}\sum_{i=1}^{n}|s_{ij}(\phi,\gamma)| \leqslant C_s < \infty$ for some positive constant C_s. The product matrix of such matrices also shares the same property. The product of $S(\phi, \gamma)$ with a matrix whose elements are uniformly bounded in absolute value has elements uniformly bounded in absolute value. See, e.g., Remark A.1 in Kelejian and Prucha (2004). The following lemma summarizes some well understood key properties of these matrices for the convenience of the reader. Proofs are straight forward and omitted, and can e.g., be found in Kuersteiner, Prucha, and Zeng (2023).

Lemma B.1. *The matrices I_c^*, J_c^* and $pI_c^* + qJ_c^*$ have the following properties:*

(i) They are symmetric, idempotent, orthogonal, $I_c^* + J_c^* = I$ and $I_c^*\mathbf{1}_c = 0$, $J_c^*\mathbf{1}_c = \mathbf{1}_c$.

(ii) $\det\left(pI_c^* + qJ_c^*\right) = p^{(n_c-1)}q$.

(iii) The eigenvalues of $pI_c^* + qJ_c^*$ are p $(n-1$ times) and q (once).

(iv) For p_1, p_2 ,q_1 and q_2 arbitrary constants, $(p_1I_c^* + q_1J_c^*)(p_2I_c^* + q_2J_c^*) = (p_1p_2I_c^* + q_1q_2J_c^*)$, and thus multiplication of matrices of the form $pI_c^* + qJ_c^*$ is commutative and associative.

(v) If $p \neq 0$, $q \neq 0$, $(pI_c^* + qJ_c^*)$ is invertible with $(pI_c^* + qJ_c^*)^{-1} = p^{-1}I_c^* + q^{-1}J_c^*$.

(vi) $\text{tr}(pI_c^* + qJ_c^*) = p(n_c - 1) + q$.

(vii) Both I_c^* and J_c^* have row and column sum uniformly bounded in absolute value. When p and q are bounded in absolute value, the row and column sums of $pI_c^* + qJ_c^*$ are uniformly bounded in absolute value.

Corollary B.1. *Suppose Assumptions 3, 5, 4, 6, 7, and 9 hold. Then $I + \rho M$, $\Omega(\gamma)$, and Σ_t are singular. If $S_a(\phi, \gamma)$ and $S_b(\phi, \gamma)$ are elements or product matrices of elements from*

$$\left\{M, (I + \rho M), (I + \rho M)^{-1}, \Omega(\gamma), \Omega(\gamma)^{-1}, \Sigma_t^{-1}, \Sigma_t\right\} \tag{38}$$

then (i) $S_a(\phi,\gamma)$ and $S_a(\phi,\gamma)A$ is measurable w.r.t $\mathcal{Z}_n^*$, continuous in ϕ and γ, and has row and column sums uniformly bounded in absolute value.

(ii) For $\Upsilon \in \{H, \mu^*, \underline{X}\}$, $S_a(\phi,\gamma)\Upsilon$, $AS_a(\phi,\gamma)\Upsilon$ and $S_aAS_b\Upsilon$ are measurable w.r.t. $\mathcal{Z}_n^*$, continuous in ϕ and γ, and have elements uniformly bounded in absolute value.

Proof. By Assumptions 3, 4 and 6, $(n_c-1-\rho)\,/\,(n_c-1)$, $(1+\rho)$, γ_j^2, ρ_{0,τ_c}^2, and $\sigma_{0,t}^2$ in (35), (36), (37) are uniformly bounded below by 0. Hence by B.1(v), $I+\rho M$, $\Omega(\gamma)$ and Σ_t are non-singular. An inspection of Equations (34)-(37) shows that all matrices in the set can be written in the form of $\operatorname{diag}_{c=1}^{C}\left(p_cI_c^*+q_cJ_c^*\right)$, with p_c and q_c measurable w.r.t $\mathcal{Z}_n^*$, continuous in ϕ and γ, and uniformly bounded. Also note that by Assumption 5, the row and column sums of A are also uniformly bounded in absolute value. Utilizing remark A.1 in Kelejian and Prucha (2004), part (i) of the corollary thus follows. Note that $\Upsilon \in \{H, \mu^*, \underline{X}\}$ is measurable w.r.t. $\mathcal{Z}_n^*$, and has elements uniformly bounded in absolute value under Assumption 7(i). Applying remark A.1 in Kelejian and Prucha (2004) again, part (ii) of the corollary holds. □

We use these properties frequently in the derivations that follow.

Lemma B.2. *Suppose Assumptions 3, 4, 5, , 6, 8 and 9 hold. Consider the block diagonal matrices $\mathcal{M}(\rho)=\operatorname{diag}_{c=1}^{C}\left(\mathcal{M}_c(\rho)\right)$ with $\mathcal{M}_c(\rho)=\left(I_c+\rho M_c\right)^{-1}\left(I_c+\rho_0M_c\right)$. Then for all $\gamma\in\Gamma$,*

$$\operatorname{tr}\left(\mathcal{M}(\rho)\,\Omega(\gamma)^{-1/2}A\Omega(\gamma)^{-1/2}\mathcal{M}(\rho)\,\Omega_0\right)=0$$

and

$$\lim_{n\to\infty}\frac{1}{n}\operatorname{tr}\left(\mathcal{M}(\rho)\,\Omega(\gamma)^{-1/2}A\Omega(\gamma)^{-1/2}\mathcal{M}(\rho)\,\Omega_0\right)=0$$

each has only one solution for $\rho\in(-1,1)$. This solution is given by $\rho=\rho_0$.

Proof. Let $D(\rho,\gamma)=\mathcal{M}(\rho)\,\Omega(\gamma)^{-1/2}A\Omega(\gamma)^{-1/2}\mathcal{M}(\rho)\,\Omega_0$ and observe that $D(\rho,\gamma)=\operatorname{diag}_{c=1}^{C}\left(D_c(\rho,\gamma)\right)$ with

$$\begin{aligned}D_c(\rho,\gamma) &= \mathcal{M}_c(\rho)\left(\gamma_{\tau_c}^2I_c\right)^{-1/2}A_c\left(\gamma_{\tau_c}^2I_c\right)^{-1/2}\mathcal{M}_c(\rho)\left(\gamma_{0,\tau_c}^2I_c\right)\\ &= \frac{\gamma_{\tau_c,0}^2}{\gamma_{\tau_c}^2}\mathcal{M}_c(\rho)\,A_c\mathcal{M}_c(\rho)\,.\end{aligned}$$

Noting that $\operatorname{tr}(A_c)=0$ and hence $\operatorname{tr}(A_cI_c^*)=\operatorname{tr}(A_c(I-J_c^*))=-\operatorname{tr}(A_cJ_c^*)$, and using that in light of Lemma B.1

$$\mathcal{M}_c(\rho) = \left(I_c+\rho M_c\right)^{-1}\left(I_c+\rho_0M_c\right)=\frac{n_c-1-\rho_0}{n_c-1-\rho}I_c^*+\frac{1+\rho_0}{1+\rho}J_c^*$$

we have

$$\begin{aligned}\operatorname{tr}\left(\mathcal{M}_c\left(\rho\right)A_c\mathcal{M}_c\left(\rho\right)\right) &= \operatorname{tr}\left(A_c\left[\left(\frac{n_c-1-\rho_0}{n_c-1-\rho}\right)^2 I_c^* + \left(\frac{1+\rho_0}{1+\rho}\right)^2 J_c^*\right]\right)\\ &= \left[\left(\frac{1+\rho_0}{1+\rho}\right)^2 - \left(\frac{n_c-1-\rho_0}{n_c-1-\rho}\right)^2\right]\operatorname{tr}\left(A_cJ_c^*\right)\\ &= (\rho_0-\rho)C(\rho,n_c)\operatorname{tr}\left(A_cJ_c^*\right)\end{aligned}$$

where

$$C(\rho,n_c) = \left[\left(\frac{1+\rho_0}{1+\rho}\right)+\left(\frac{n_c-1-\rho_0}{n_c-1-\rho}\right)\right]\left[\frac{1}{1+\rho}+\frac{1}{n_c-1-\rho}\right].$$

Hence

$$\operatorname{tr}\left(D(\rho,\gamma)\right) = \sum_{c=1}^{C}\operatorname{tr}\left(D_c(\rho,\gamma)\right) = (\rho_0-\rho)\sum_{c=1}^{C}\frac{\gamma_{\tau_c,0}^2}{\gamma_{\tau_c}^2}C(\rho,n_c)\operatorname{tr}\left(A_cJ_c^*\right).$$

Clearly, at $\rho=\rho_0$, for all $\gamma\in\Gamma$, we have $\operatorname{tr}\left(D(\rho_0,\gamma)\right)=0$ for all n.

Next observe that $\operatorname{tr}\left(A_cJ_c^*\right) = \mathbf{1}_c^{\prime}A_c\mathbf{1}_c/n_c > 0$ and that for any $n_c \geqslant 2$, and $-1<\rho<1$, both $n_c-1-\rho$ and $1+\rho$ are positive, and therefore $C(\rho,n_c)$ is positive for all $\rho\in(-1,1)$. Therefore, $\operatorname{tr}\left[\mathcal{M}_c\left(\rho\right)A_c\mathcal{M}_c\left(\rho\right)\right]<0$ if $\rho>\rho_0$, $\operatorname{tr}\left[\mathcal{M}_c\left(\rho\right)A_c\mathcal{M}_c\left(\rho\right)\right]>0$ if $\rho<\rho_0$ and $\operatorname{tr}\left[\mathcal{M}_c\left(\rho\right)A_c\mathcal{M}_c\left(\rho\right)\right]$ has the same sign for all c. In all, for any value of γ the only solution for

$$\operatorname{tr}\left(\mathcal{M}\left(\rho\right)\Omega(\gamma,\theta)^{-1/2}A\Omega(\gamma,\theta)^{-1/2}\mathcal{M}\left(\rho\right)\Omega_0\right)=0$$

is $\rho=\rho_0$. Fruthermore, by Assumptions 4 and 6 there exists some constants c_D and c_γ such that for any ρ_0 in the interior of the parameter space, $C(\rho,n_c)>C_D$ and $\frac{\gamma_{\tau_c,0}^2}{\gamma_{\tau_c}^2}>c_\gamma>0$. This implies that,

$$|\operatorname{tr}(D(\rho))| \geqslant |\rho_0-\rho|\sum_{c=1}^{C}C_Ac_Dc_\varsigma,$$

$$\left|\frac{1}{n}\operatorname{tr}(D(\rho))\right| \geqslant |\rho_0-\rho|C_Ac_Dc_\varsigma\frac{C}{n}.$$

By Assumption 6, $C/n \geq 1/\bar{n}_c > 0$ such that the conclusion follows. □

Lemma B.3. *Let Assumption 2 hold. Then,* u_{it} *defined in Theorem 2.1 is a martingale difference sequence w.r.t to the filtration* $\mathcal{A}_{n,\nu}$ *with* $\mathcal{A}_{n,0}=\mathcal{Z}_n^*=\sigma\left(S,\alpha,X,z,\zeta^*\right)$ and

$$\mathcal{A}_{n,(t-1)n+i-1}=\sigma\left(S,\alpha,X,z,\left\{y_{j0}^*,u_{j1}...,u_{j,t-1}\right\}_{j=1}^{n},\left\{u_{jt}\right\}_{j=1}^{i-1}\right)\ \text{for}\ t=1,2;i=1,\ldots,n. \qquad (39)$$

In addition, $\sup_{i,t}E\left[|u_{it}|^{4+\eta}\right]<\infty$ *for some* $\eta>0$.

Proof. First note that the filtrations $\mathcal{A}_{n,\nu}$ are increasing in the sense $\mathcal{A}_{n,0}\subseteq\mathcal{A}_{n,1}....\subseteq\mathcal{A}_{n,\nu}\subseteq\mathcal{A}_{n,\nu+1}.....$ By construction $u_{it}=y_{it}^*-\kappa_if_t$ where κ_if_t is $\mathcal{A}_{n,0}$-measurable. For general i and

t note that since $u_{i,t} = y^*_{i,t} - \kappa_i f_t$ and $\kappa_i f_t$ is $\mathcal{Z}^*_n$-measurable it follows that $\mathcal{B}_{n,i,t} \subseteq \mathcal{F}_{n,i,t}$. Since $y^*_{i,t} = \kappa_i f_t + u_{i,t}$ and $\kappa_i f_t$ is $\mathcal{Z}^*_n$-measurable it follows that $\mathcal{F}_{n,i,t} \subseteq \mathcal{B}_{n,i,t}$. Consequently $\mathcal{F}_{n,i,t} = \mathcal{B}_{n,i,t}$. Observing that $\mathcal{A}_{n,(t-1)n+i-1} \subseteq \mathcal{B}_{n,i,t}$ it follows, using iterated expectations, that

$$E\left[u_{it}|\mathcal{A}_{n,(t-1)n+i-1}\right] = E\left[E\left[u_{it}|\mathcal{B}_{n,i,t}\right]|\mathcal{A}_{n,(t-1)n+i-1}\right] = 0$$

observing that $E\left[u_{it}|\mathcal{B}_{n,i,t}\right] = E\left[y^*_{i,t} - \kappa_i f_t|\mathcal{F}_{n,i,t}\right] = 0$ in light of Assumption 2. Finally observe that u_{it} is measurable w.r.t. $\mathcal{A}_{n,(t-1)n+i}$.

Recall that by Assumption 2 there exists a random variable y such that $|y^*_{it}| + |\kappa_i f_t| \leq y$ with $E\left[|y|^{4+\eta}|\mathcal{A}_{n,0}\right] \leq K_y < \infty$. Hence

$$E\left[|u_{it}|^{4+\eta}|\mathcal{A}_{n,0}\right] \leq E\left[(|y^*_{it}| + |\kappa_i f_t|)^{4+\eta}|\mathcal{A}_{n,0}\right] \leq E\left[|y|^{4+\eta}|\mathcal{A}_{n,0}\right] < K_y < \infty. \tag{40}$$

Since the bound on the RHS is uniform in i and t by assumption it follows that $\sup_{i,t} E\left[|u_{it}|^{4+\eta}\right] < \infty$. □

C Proofs

Proof of Theorem 2.1. By the assumption on the boundedness of the $4+\eta$ moments all considered expectations and conditional expectations exist. Since $u_{it} = y^*_{it} - E\left[y^*_{it}|\mathcal{F}_{n,i,t}\right]$ we have $E\left[u_{it}|\mathcal{F}_{n,i,t}\right] = 0$ by construction. Observing that $E\left[y^*_{it}|\mathcal{F}_{n,i,t}\right] = \kappa_i f_t$ and that κ_i is $\mathcal{Z}^*_n$-measurable it follows that conditional on $\mathcal{Z}^*_n$ knowledge if y^*_{it} is equivalent to knowledge of u_{it}. Given that $\mathcal{Z}^*_n \subset \mathcal{F}_{n,i,t}$ it follows that $\mathcal{F}_{n,i,t} = \sigma\left(S, \alpha, X, \left\{y^*_{j,0}, ..., y^*_{j,t-1}\right\}_{j=1}^n, y^*_{-i,t}\right)$ for $t = 1, 2$ can be written equivalently as

$$\begin{aligned}
\mathcal{F}_{n,i,1} &= \sigma\left(S, \alpha, X, \zeta^*, u_{-i,1}\right) \\
\mathcal{F}_{n,i,2} &= \sigma\left(S, \alpha, X, \zeta^*, \{u_{i1}\}_{j=1}^n, u_{-i,2}\right)
\end{aligned}$$

with $u_{-i,t} = [u_{1t}, .., u_{i-1,i}, u_{i+1,t}, ..., u_{n,t}]'$. Next observe that the u_{jt} are $\mathcal{F}_{n,i,t}$-measurable for $i \neq j$, and hence by iterated expectations $\operatorname{Cov}\left(u_{it}u_{jt}|\mathcal{F}_{n,i,t}\right) = E\left[u_{jt}E\left[u_{it}|\mathcal{F}_{n,i,t}\right]\right] = 0$. Observe further that for $s < t$ the u_{js} are $\mathcal{F}_{n,i,t}$-measurable for all i, and hence by iterated expectations $\operatorname{Cov}\left(u_{it}u_{js}|\mathcal{F}_{n,i,t}\right) = E\left[u_{js}E\left[u_{it}|\mathcal{F}_{n,i,t}\right]\right] = 0$. The corresponding claims regarding the mean and covariances of the u_{it} conditional on $\mathcal{G}_{n,i,t} \subseteq \mathcal{F}_{n,i,t}$ follow immediately by iterated expectations. This proves the first part of the lemma.

To prove the second part, let $\mathcal{G}_{n,i,t} \subseteq \mathcal{F}_{n,i,t}$ be some information set. Then given Assumption 2 holds, it follows by iterated expectations that

$$\mathrm{E}[y^*_{ii} - \kappa_i f_t|\mathcal{G}_{n,i,t}] = E\left[E\left[y^*_{it}|\mathcal{F}_{n,i,t}\right]|\mathcal{G}_{n,i,t}\right] - E\left[\kappa_i f_t|\mathcal{G}_{n,i,t}\right] = E\left[\kappa_i f_t|\mathcal{G}_{n,i,t}\right] - E\left[\kappa_i f_t|\mathcal{G}_{n,i,t}\right] = 0.$$

Thus Assumption 2 cannot hold if $E\left[y^*_{ii} - \kappa_i f_t | \mathcal{G}_{n,i,t}\right] \neq 0$. $\square$

Note that Lemma 4.1, Corollary 4.1, and Lemma 4.3 can be viewed as special cases of Lemma 4.2. We therefore prove Lemma 4.2 first.

Proof of Lemma 4.2. In light of (5) it follows that,

$$y_1 - f_1 y_2 = \mu^*(f_{0,1} - f_1) + \underline{X}\delta(f_1, \rho_0, \beta_0) + (I + \rho_0 M)(u_1 - f_1 u_2), \tag{41}$$

where $\delta(f_1, \rho_0, \beta_0)$ is defined in (7) and $\delta(f_{1,0}, \rho_0, \beta_0) = \delta_0$.

Substitution of (41) into (13) yields

$$\begin{aligned} u^+(\phi,\gamma) &= \Omega(\gamma)^{-1/2}(I+\rho M)^{-1} \\ &\quad \times \left\{ (f_{0,1} - f_1)\,\mu^* + \underline{X}\left(\delta(f_1,\rho_0,\beta_0) - \delta\right) + (I + \rho_0 M)(u_1 - f_1 u_2) \right\} \qquad (42) \\ &= V(\rho,\gamma) \begin{pmatrix} \underline{X}, & \mu^* \end{pmatrix} \begin{pmatrix} \delta(f_1,\rho_0,\beta_0) - \delta \\ f_{0,1} - f_1 \end{pmatrix} + \Omega(\gamma)^{-1/2}\mathcal{M}(\rho)(u_1 - f_1 u_2), \qquad (43) \end{aligned}$$

recalling that $V(\rho,\gamma) = \Omega(\gamma)^{-1/2}(I+\rho M)^{-1}$ and $\mathcal{M}(\rho) = (I+\rho M)^{-1}(I+\rho_0 M)$.

We first analyze the linear moment conditions. Note that

$$H'u^+(\phi,\gamma) = H'V(\rho,\gamma) \begin{pmatrix} \underline{X}, & \mu^* \end{pmatrix} \begin{pmatrix} \delta(f_1,\rho_0,\beta_0) - \delta \\ f_{0,1} - f_1 \end{pmatrix} + H'\Omega(\gamma)^{-1/2}\mathcal{M}(\rho)(u_1 - f_1 u_2). \tag{44}$$

Let

$$K_n(\rho,\gamma) = n^{-1}H'V(\rho,\gamma)\ (\underline{X}, \quad \mu^*) \tag{45}$$

Observe that

$$n^{-1}E\left[H'u^+(\phi,\gamma)\,|\mathcal{Z}_n\right] = E\left(K_n(\rho,\gamma)\,|\mathcal{Z}_n\right) \begin{pmatrix} \delta(f_1,\rho_0,\beta_0) - \delta \\ f_{0,1} - f_1 \end{pmatrix}. \tag{46}$$

where the r.h.s. is seen to hold in light of (44) and (45), and since H is $\mathcal{Z}_n$-measurable and thus

$$E\left[H'\Omega(\gamma)^{-1/2}\mathcal{M}(\rho)(u_1 - f_1 u_2)\,|\mathcal{Z}_n\right] = 0.$$

The assumptions imply that K_n has full rank a.s. To see this, note that in light of (35), (36), Lemma (B.1)(iv) implies that

$$V(\rho,\gamma) = \mathrm{diag}_{c=1}^{C}\left(\frac{1}{\gamma_{\tau_c}}\frac{n_c - 1 - \rho}{n_c - 1}I^*_c + \frac{1}{\gamma_{\tau_c}}\left(1+\rho\right)J^*_c\right). \tag{47}$$

Under Assumptions 4 and 6, there exists some constants c_e and C_e such that $0 < c_e < \frac{1}{\gamma_{\tau_c}} \frac{n_c - 1 - \rho}{n_c - 1} < C_e < \infty$ and $0 < c_e < \frac{1}{\gamma_{\tau_c}}(1+\rho) < C_e < \infty$. In light of Lemma B.1(iii), the eigenvalues of $V(\rho, \gamma)$ are uniformly bounded below by $c_e > 0$. By Assumption 7 we have $\lambda_{min}\left(n^{-1}\underline{X}'\underline{X}\right) \geq \xi_X > 0$ and thus $\lambda_{min}\left(n^{-1}\underline{X}'V(\rho,\gamma)\underline{X}\right) \geq c_e \xi_X > 0$ uniformly for all γ, ρ, and n.

Recalling that $H = (\underline{X}, z)$ and recalling the expression for $K_n(\theta, \gamma)$ given in (45) the matrix can be written as

$$K_n(\rho,\gamma) = \frac{1}{n}\begin{pmatrix} \underline{X}'V(\rho,\gamma)\underline{X} & \underline{X}'V(\rho,\gamma)\mu^* \\ z'V(\rho,\gamma)\underline{X} & z'V(\rho,\gamma)\mu^* \end{pmatrix}. \tag{48}$$

Thus by the determinant for partitioned matrices, and using the shorthand notation $V = V(\rho, \gamma)$,

$$|\det\left(E\left[K_n(\rho,\gamma)\,|\mathcal{Z}_n\right]\right)| = \left|\det\left(\frac{1}{n}E\left[\underline{X}'V\underline{X}\,|\mathcal{Z}_n\right]\right)\right|\left|\det\left\{E\left[\left(\frac{1}{n}z'V\mu^* - \frac{1}{n}z'V\underline{X}(\frac{1}{n}\underline{X}'V\underline{X})^{-1}\frac{1}{n}\underline{X}'V\mu^*\right)|\mathcal{Z}_n\right]\right\}\right| \tag{49}$$

$$\geq \xi_X c_e \left|\frac{1}{n}E\left[(V^{1/2}z)'Q_{V^{1/2}X}(V^{1/2}\mu^*)\,|\mathcal{Z}_n\right]\right|.$$

Observe that $\mathcal{Z}_n \subseteq \mathcal{Z}_n^*$, and that V, $\underline{X}$ and μ^* are measurable w.r.t $\mathcal{Z}_n^*$, so

$$E\left[(V^{1/2}z)'Q_{V^{1/2}X}(V^{1/2}y_2)\,|\mathcal{Z}_n\right] \quad = \quad E\left[(V^{1/2}z)'Q_{V^{1/2}X}(V^{1/2}\mu^*)\,|\mathcal{Z}_n\right]$$

in light of (8) and since $Q_{V^{1/2}X}$ is orthogonal to $V^{1/2}\underline{X}$ and $E\left[u_{i2}|Z_n\right] = 0$ by Theorem 2.1. Thus

$$\begin{aligned}\inf_{\gamma\in\Gamma,\rho\in[-K_\rho,K_\rho]}|\det\left(E\left[K_n(\rho,\gamma)\,|\mathcal{Z}_n\right]\right)| \quad &\geq \quad \xi_X c_e \inf_{\gamma\in\Gamma,\rho\in[-K_\rho,K_\rho]} n^{-1}\left|E\left[(V^{1/2}z)'Q_{V^{1/2}X}(V^{1/2}y_2)\,|\mathcal{Z}_n\right]\right| \\ &\geq \quad \xi_X c_e K_y > 0\end{aligned} \tag{50}$$

in light of Assumption 7. This proves that $E\left[K_n(\rho,\gamma)\,|\mathcal{Z}_n\right]$ has full rank for all admissible values of ρ and γ, and consequently the moment condition (46) equates to zero if and only if $f_1 = f_{1,0}$ and $\delta = \delta_0$.This means that $E\left[n^{-1/2}m_n^{(l)}(\phi,\gamma)\right] = 0$ if and only if $f_1 = f_{1,0}$ and $\delta = \delta_0$ for all admissible ρ and γ. Therefore $f_{1,0}$ and δ_0 are identified from the linear moment condition.

Once $f_{1,0}$ and δ_0 are identified, ρ can be identified from the quadratic moment condition. To see this, note that when evaluated at δ_0 and $f_{1,0}$, $u_1^+(\phi,\gamma)$ in Equation (42) becomes

$$\begin{aligned} u^+(\phi,\gamma)|_{f_{1,0},\delta_0} &= \Omega(\gamma)^{-1/2}(I+\rho M)^{-1}(I+\rho_0 M)(u_1 - f_{1,0}u_2) \\ &= \Omega(\gamma)^{-1/2}\mathcal{M}(\rho)(u_1 - f_{1,0}u_2)\end{aligned}$$

Then,

$$
\begin{aligned}
&n^{-1}\, E\left[u^{+}(\phi,\gamma)' A u^{+}(\phi,\gamma)|\mathcal{Z}_n\right]\Big|_{f_{1,0},\delta_0} \\
=&n^{-1} E\left[(u_1 - f_{0,1}u_2)'\mathcal{M}\left(\rho\right)\Omega(\gamma)^{-1/2} A \Omega(\gamma)^{-1/2}\mathcal{M}\left(\rho\right)(u_1 - f_{0,1}u_2)|\mathcal{Z}_n\right] \\
=&n^{-1} E\left[\operatorname{tr}\left(\mathcal{M}\left(\rho\right)\Omega(\gamma)^{-1/2} A \Omega(\gamma)^{-1/2}\mathcal{M}\left(\rho\right)\Omega_0\right)|\mathcal{Z}_n\right].
\end{aligned}
$$

By Lemma B.2 the equation above equals 0 if and only if $\rho = \rho_0$ for $-1 < \rho < 1$. This establishes that for all admissible γ we have $E\left[n^{-1/2} m_n\left(\phi_0,\gamma\right)|\mathcal{Z}_n\right] = 0$ a.s. and $E\left[n^{-1/2} m_n\left(\phi,\gamma\right)|\mathcal{Z}_n\right] \neq 0$ a.s. for $\phi \neq \phi_0$. □

Proof of Lemma 4.1 and Corollary 4.1. Lemma 4.1 is a special case of Lemma 4.2 without covariates.

Note that Assumptions 1, 2, 3, 4, 5, and 6 hold. Without covariatesEquation (48) reduces to $K_n\left(\rho,\gamma\right) = z'V\left(\rho,\gamma\right)\mu^{*}$ where $V\left(\rho,\gamma\right) = \Omega\left(\gamma\right)^{-1/2}\left(I + \rho M\right)^{-1}$.

Recall that $z = [z_1', ..., z_c', ..., z_C']$ with $z_c = \grave{z}_c 1_c$. Utilizing Lemma B.1,

$$
(I_c + \rho M_c)^{-1} z_c = \left(\frac{n_c - 1 - \rho}{n_c - 1} I_c^{*} + \frac{1}{1+\rho} J_c^{*}\right)\grave{z}_c \mathbf{1}_c = \frac{1}{1+\rho}\grave{z}_c \mathbf{1}_c = \frac{z_c}{1+\rho}
$$

leads to

$$
n^{-1} E\left[y_2'\Omega(\gamma)^{-1/2}(I + \rho M)^{-1} z|\mathcal{Z}_n\right] = \frac{1}{1+\rho} n^{-1} E\left[y_2'\Omega(\gamma)^{-1/2} z|\mathcal{Z}_n\right] \geq \frac{1}{1+\rho} K_y > 0
$$

where the inequality follows from the conditions imposed in the Lemma, in particular (24). With $\rho \in (-1,1)$,E$[K_n\left(\rho,\gamma\right)|\mathcal{Z}_n] \neq 0$ such that (46) only has one solution, $f_1 = f_{1,0}$. Identification of ρ_0 from the quadratic moment is not affected by the absence of covariates. Thus Lemma 4.1 follows from Lemma 4.2.

To prove Corollary 4.1 observe that when $z = 1_n$, $n^{-1}\sum_{c=1}^{C} E\left[y_{c2}'\mathbf{1}_c\right] > K_y > 0$ implies that (24) holds, observing that

$$
n^{-1}\left|E\left[y_2'\Omega(\gamma)^{-1/2} 1_n\,|\mathcal{Z}_n\right]\right| = n^{-1} E\sum_{c=1}^{C}\left[\frac{y_{c2}'\mathbf{1}_c}{\gamma_{\tau_c}}\,|\mathcal{Z}_n\right] \geq \frac{1}{K_\gamma}\frac{1}{n}\sum_{c=1}^{C} E\left[y_{c2}'\mathbf{1}_c\,|\mathcal{Z}_n\right] \geq \frac{K_y}{K_\gamma} > 0.
$$

□

Proof of Lemma 4.3. Observe that utilizing (8) we have

$$
\begin{aligned}
n^{-1} E\left[H'(I + \rho M)\epsilon^{+}\left(\rho, f_1, \delta\right)|\mathcal{Z}_n\right] &= n^{-1} E\left[H'\left(y_1 - f_1 y_2 - \underline{X}\delta\right)|\mathcal{Z}_n\right] \\
&= n^{-1} E\left[H'\left(\mu_{*}(f_{0,1} - f_1) + \underline{X}\delta(f_1, \rho_0, \beta_0)\right)|\mathcal{Z}_n\right].
\end{aligned}
$$

Let $\tilde{K}_n = n^{-1}H'(\underline{X}, \mu^*)$, then the linear moment condition can be expressed as

$$n^{-1}E\left[H'(I+\rho M)\epsilon^{+}(\rho, f_1, \delta)\,|\mathcal{Z}_n\right] = E\left[\tilde{K}_n\,|\mathcal{Z}_n\right]\begin{pmatrix} \delta(f_1, \rho_0, \beta_0) - \delta \\ f_{0,1} - f_1 \end{pmatrix} = 0 \qquad (51)$$

Observe that $E\left[\tilde{K}_n\right]$ as full rank, as is readily seen from the proof of Lemma 4.2 in light of (50) with $V = I$. This implies that the only solution to (51) is $\delta = \delta_0$ and $f_1 = f_{1,0}$. It now follows that

$$\varepsilon^{+}(\rho, f_{0,1}, \delta_0) = \mathcal{M}(\rho)(u_1 - f_{0,1}u_2)$$

such that the quadratic moment condition is

$$E\left[\varepsilon^{+}(\rho, f_{1,0}, \delta_0)' A \varepsilon^{+}(\rho, f_{0,1}, \delta_0)\,|\mathcal{Z}_n\right] = \operatorname{tr}\left(\mathcal{M}(\rho) A \mathcal{M}(\rho)\Omega_0\right) = 0. \qquad (52)$$

By Lemma B.2 it follows that the only solution to (52) is $\rho = \rho_0$. □

Proof of Theorem 4.1. To prove part (i) of the theorem we establish the conditions of Lemma 3.1 in Pötscher and Prucha (1997).

For the linear moment conditions use (44). We have

$$\begin{aligned} n^{-1}H'u_1^{+}(\phi,\gamma) &= (f_{1,0} - f_1)\, n^{-1}H'V(\rho,\gamma)\mu^* \\ &+ \frac{1}{n}H'V(\rho,\gamma)\underline{X}\left(\delta(f_1,\rho_0,\beta_0) - \delta\right) \\ &+ \frac{1}{n}H'\Omega(\gamma)^{-1/2}\mathcal{M}(\rho)(u_1 - f_1u_2). \\ &\equiv I + II + III. \end{aligned}$$

Consider I and II. Using Assumption 8 and noting that $V(\rho,\gamma)$ defined in (47) satisfies the assumption on $C(\phi,\gamma)$, hence $n^{-1}H'V(\rho,\gamma)\underline{X}$ converges uniformly to

$$\lim_{n\to\infty} E\left[n^{-1}H'V(\rho,\gamma)\underline{X}\right] \equiv \mathscr{U}_{H,x}(\phi,\gamma) \text{ a.s.} \qquad (53)$$

and $n^{-1}H'V(\rho,\gamma)\mu^*$ converges uniformly to

$$\lim_{n\to\infty} E\left[n^{-1}H'V(\rho,\gamma)\mu^*\right] \equiv \mathscr{U}_{H,\mu}(\phi,\gamma) \text{ a.s..} \qquad (54)$$

Consider III. From Corollary B.1 and Assumption 8, we have $sup_{\phi,\gamma}|III| \to_p 0$.

Now consider the quadratic moment conditions. Let $\Upsilon_1 = [\underline{X}, \mu^*]$, $\vartheta = (\delta', f_1)'$, $\vartheta_0 = (\delta(f_1,\rho_0,\beta_0)', f_{0,1})'$, then in light of (42) we have

$$\begin{aligned} n^{-1}u_1^{+}(\phi,\gamma)Au_1^{+}(\phi,\gamma) &= n^{-1}\left(\vartheta-\vartheta_0\right)'\Upsilon_1'V(\rho,\gamma)'AV(\rho,\gamma)\Upsilon_1\left(\vartheta-\vartheta_0\right) \\ &\quad +n^{-1}2\left(\vartheta-\vartheta_0\right)'\Upsilon_1'V(\rho,\gamma)'A\Omega(\gamma)^{-1/2}\mathcal{M}(\rho)\left(u_1-f_1u_2\right) \\ &\quad +n^{-1}\left(u_1-f_1u_2\right)'\mathcal{M}(\rho)'\Omega(\gamma)^{-1/2}A\Omega(\gamma)^{-1/2}\mathcal{M}(\rho)\left(u_1-f_1u_2\right) \\ &= I+II+III. \end{aligned} \tag{55}$$

By Assumption 8(i) and (ii) , $n^{-1}\Upsilon_1'V(\rho,\gamma)'AV(\rho,\gamma)\Upsilon_1$ in I converges uniformly to

$$\lim_{n\to\infty} n^{-1}E\left[\Upsilon_1'V(\rho,\gamma)'AV(\rho,\gamma)\Upsilon_1\right] \equiv \mathscr{U}_{A,\Upsilon}\left(\phi,\gamma\right) \tag{56}$$

By Corollary B.1 and Assumption 8, II converges uniformly to 0. Consider III, let $B(\phi,\gamma)=\mathcal{M}(\rho)'\Omega(\gamma)^{-1/2}A\Omega(\gamma)^{-1/2}\mathcal{M}(\rho)$, then

$$\frac{1}{n}(u_1-f_1u_2)'B(\phi,\gamma)(u_1-f_1u_2)=\frac{1}{n}u_1'B(\phi,\gamma)u_1+\frac{f_1^2}{n}u_2'B(\phi,\gamma)u_2-\frac{2}{n}f_1u_1'B(\phi,\gamma)u_2.$$

Each term on the R.H.S. converges to its mean uniformly in ϕ and γ under Corollary (B.1) and Assumption 8. Consequently, $\frac{1}{n}(u_1-f_1u_2)'B(\phi,\gamma)(u_1-f_1u_2)$ converges uniformly to the limit of its mean

$$\begin{aligned} E\left[\frac{1}{n}(u_1-f_1u_2)'B(\phi,\gamma)(u_1-f_1u_2)\right] &= \operatorname{tr}\left[\frac{1}{n}B(\phi,\gamma)\left(\Sigma_1+f_1^2\Sigma_2\right)\right] \\ &= \frac{1}{n}\operatorname{tr}\left(B(\phi,\gamma)\Omega_0\right)+\frac{1}{n}\left(f_1^2-f_{0,1}^2\right)\operatorname{tr}\left(B(\phi,\gamma)\Sigma_2\right), \end{aligned}$$

observing that $\Omega_0=\Sigma_1+f_{0,1}^2\Sigma_2$.

Note that both $\operatorname{tr}\left(B(\phi,\gamma)\Omega_0\right)$ and $\operatorname{tr}\left(B(\phi,\gamma)\Sigma_2\right)$ can be written in the form of $\operatorname{tr}\left(AC(\phi,\gamma)\right)$, with $C(\phi,\gamma)$ satisfy the conditions in Assumption 8, consequently, there exist bounded and continuously differentiable functions in ϕ and γ denoted by $\mathscr{U}_{A,\Omega}(\phi,\gamma)$ and $\mathscr{U}_{A,\Sigma}(\phi,\gamma)$ such that

$$\lim_{n\to\infty}\sup_{\phi,\gamma}\left\|E\left[\frac{1}{n}\operatorname{tr}\left(B(\phi,\gamma)\Omega_0\right)\right]-\mathscr{U}_{A,\Omega}(\phi,\gamma)\right\|=0 \tag{57}$$

and

$$\lim_{n\to\infty}\sup_{\phi,\gamma}\left\|E\left[\frac{1}{n}\operatorname{tr}\left(B(\phi,\gamma)\Sigma_2\right)\right]-\mathscr{U}_{A,\Sigma}(\phi,\gamma)\right\|=0 \tag{58}$$

From uniform convergence and (53), (54), (56), (57) and (58) it follows that

$$\lim_{n\to\infty} n^{-1/2} E\left[m_n(\phi,\gamma)\right] = \left[\begin{array}{c} \left(\mathscr{U}_{H,x}(\phi,\gamma), \mathscr{U}_{H,\mu}(\phi,\gamma)\right)\begin{pmatrix} \delta(f_1,\rho_0,\beta_0)-\delta \\ f_{1,0}-f_1 \end{pmatrix} \\ (\vartheta-\vartheta_0)' \mathscr{U}_{A,\Upsilon}(\phi,\gamma)(\vartheta-\vartheta_0) + \mathscr{U}_{A,\Omega}(\phi,\gamma) + \left(f_1^2 - f_{1,0}^2\right)\mathscr{U}_{A,\Sigma}(\phi,\gamma) \end{array}\right] \tag{59}$$

$$\equiv \left[\begin{array}{c} \mathfrak{m}_l(\phi,\gamma) \\ \mathfrak{m}_q(\phi,\gamma) \end{array}\right] \equiv \mathfrak{m}(\phi,\gamma)$$

By uniform convergence of $n^{-1/2} m_n(\phi,\gamma)$ it also follows that for $\bar{\gamma}_n \to \gamma_*$

$$\sup_{\phi} \left\| n^{-1/2} m_n(\phi,\bar{\gamma}_n) - \mathfrak{m}(\phi,\gamma_*) \right\| \to 0 i.p.$$

We proceed to show, building on our results for finite n, that ϕ_0 is also the unique solution vector of the limiting moment condition $\mathfrak{m}(\phi,\gamma_*)=0$. Observe that from (45), (53) and (54) we have $\left[\left[\mathscr{U}_{H,x}(\phi,\gamma_*), \mathscr{U}_{H,\mu}(\phi,\gamma_*)\right]\right] = \lim_{n\to\infty} E\left[K_n(\rho,\gamma_*)\right]$,and by (50) we have $\inf_{\gamma\in\Gamma,\rho} \left|\det\left(E\left[K_n(\rho,\gamma)\right]\right)\right| \geq C_K > 0$ for some $C_K > 0$ and all n. Since det(.) is a continuous function it follows that for any ϕ, γ_* we have

$$\begin{aligned} \left|\det\left(\left[\left[\mathscr{U}_{H,x}(\phi,\gamma_*), \mathscr{U}_{H,\mu}(\phi,\gamma_*)\right]\right]\right)\right| &= \left|\det\left(\lim_{n\to\infty} E\left[K_n(\rho,\gamma_*)\right]\right)\right| \\ &= \lim_{n\to\infty} \left|\det\left(E\left[K_n(\rho,\gamma_*)\right]\right)\right| \geq C_K > 0. \end{aligned}$$

In light of this for any ρ and γ_*

$$\mathfrak{m}_l(\phi,\gamma_*) = \mathfrak{m}_l(\rho, f_1, \delta, \gamma_*) = 0$$

if and only if $\delta = \delta_0 = \delta(f_{1,0},\rho_0,\beta_0)$ and $f_1 = f_{1,0}$, i.e., $\vartheta = \vartheta_0$.

To show that $\mathfrak{m}(\phi,\gamma_*)=0$ if and only if $\phi = (\rho, f_1, \delta')' = \phi_0 = (\rho_0, f_{1,0}, \delta_0')'$ it thus suffices to show that $\mathfrak{m}_q(\rho, f_{1,0}, \delta_0, \gamma_*) = 0$ if and only if $\rho = \rho_0$.

Recall that

$$\begin{aligned} \mathfrak{m}_q(\rho, f_{1,0}, \delta_0, \gamma_*) &= \mathscr{U}_{A,\Omega}(\rho, f_{1,0}, \delta_0, \gamma_*) \\ &= \lim_{n\to\infty} \frac{1}{n} \operatorname{tr}\left(\mathcal{M}(\rho)\,\Omega(\gamma)^{-1/2} A \Omega(\gamma)^{-1/2} \mathcal{M}(\rho)\,\Omega_0\right). \end{aligned}$$

By Lemma B.2, $\mathfrak{m}_q(\rho, f_{0,1}, \delta_0, \gamma_*) = 0$ if and only if $\rho = \rho_0$, and consequently $\mathfrak{m}(\phi,\gamma_*)=0$ if and only if $\phi = \phi_0$. Assumption 8 implies that $\mathfrak{m}(\phi,\gamma_*)$ and thus $\mathfrak{m}(\phi,\gamma_*)' \Xi^{-1} \mathfrak{m}(\phi,\gamma_*)$ is continuous. Since the parameter space is compact it follows that $\phi_0 = [\rho_0, f_{0,1}, \delta_0']'$ is the identifiable unique minimizer of $\mathfrak{m}(\phi,\gamma_*)' \Xi^{-1} \mathfrak{m}(\phi,\gamma_*)$; cp. the discussion after Definition 3.1 in Pötscher and Prucha

(1997). Consistency now follows immediately from Lemma 3.1 in Pötscher and Prucha (1997).

For part (ii) of the theorem we first consider $\tilde{f}_1$. Observing that in light of (8) we have $y_1 = f_{1,0}y_2 + X\delta_0 + (I+\rho_0 M)(u_1 - f_{1,0}u_2)$ and that $Q_X\underline{X} = 0$ it follows from (17) that

$$\tilde{f}_1 = \frac{z'Q_X y_1}{z'Q_X y_2} = f_{1,0} + \frac{n^{-1}z'Q_X(I+\rho_0 M)(u_1 - f_{1,0}u_2)}{n^{-1}z'Q_X y_2}. \tag{60}$$

Given that the elements of $\underline{X}$ and z are uniformly bounded by Assumption 7 and observing that the row and columns sums of M are uniformly bounded in absolute value it follows, e.g., from Remark A.1 in Kelejian and Prucha (2004) that the elements of $z'Q_X(I+\rho_0 M)$ are uniformly bounded. Thus $n^{-1}z'Q_X(I+\rho_0 M)(u_1 - f_{1,0}u_2) \rightarrow_p 0$ since $n^{-1}z'Q_X(I+\rho_0 M)u_t \rightarrow_p 0$ for $t = 1, 2$ by Assumption 8.

Observe that for $\rho = 0$, $f_1 = 1$, $\rho^2_{\tau_c} = 1$, $\sigma^2 = 1$, $V(\gamma,\theta) = I_n$ and $Q_{V^{1/2}X} = Q_X$. Using (5) implies that

$$n^{-1}z'Q_X y_2 = n^{-1}z'Q_X\mu^* + n^{-1}z'Q_X(I+\rho_0 M)u_2$$

and $E\left[n^{-1}z'Q_X y_2|\mathcal{Z}_n\right] = E\left[n^{-1}z'Q_X\mu^*|\mathcal{Z}_n\right]$. As argued above, $n^{-1}z'Q_X(I+\rho_0 M)u_2 \rightarrow_p 0$. By Assumption 8(i) and (ii) it follows that $n^{-1}\left|z'Q_X\mu^* - E\left[z'Q_X\mu^*\right]\right| \rightarrow 0$ a.s. and $n^{-1}\left|E\left[z'Q_X\mu^*\right] - \mathscr{U}_{z\mu}\right| \rightarrow 0$ a.s. Assumption 7(iii) implies that $n^{-1}\left|E\left[z'Q_X\mu^*\right]\right| \geq K_y > 0$ which in turn implies that $|\mathscr{U}_{z\mu}| \geq K_y > 0$. This shows that the denominator on the r.h.s. of (60) converges to a positive constant. Having shown that the numerator converges to zero it follows that $\tilde{f}_1 \rightarrow_p f_{1,0}$.

Then, turning to

$$\begin{aligned}
\tilde{\delta} &= \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\left(y_1 - y_2\tilde{f}_1\right) \\
&= \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\left(\underline{X}\delta_0 + (I_n + \rho_0 M)(u_1 - u_2 f_{1,0})\right) \\
&\quad + \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'y_2\left(f_{1,0} - \tilde{f}_1\right) \\
&\equiv I + II
\end{aligned}$$

where we used (8) for $y_1 - f_{1,0}y_2$. By Assumptions 7 and 8 it follows that $n^{-1}\underline{X}'\underline{X} \rightarrow_p \mathscr{U}_{XX}$ where $\mathscr{U}_{XX}$ is positive definite and symmetric and $n^{-1}\underline{X}'y_2 \rightarrow_p \mathscr{U}_{Xy}$ where$\mathscr{U}_{Xy}$ is bounded. This implies that $II \rightarrow_p 0$. For I recall from above that $n^{-1}\underline{X}'(I_n + \rho_0 M)(u_1 - u_2 f_{0,1}) \rightarrow_p 0$. This shows that $\tilde{\delta} \rightarrow_p \delta_0$.

Now consider the quadratic moment conditions. Let $\Upsilon_1 = [X, y_2]$, $\tilde{\vartheta} = \left(\tilde{\delta}, \tilde{f}_1\right)$, and observe that utilizing (8) we have

$$\begin{aligned}
\epsilon^+\left(\rho, \tilde{f}_1, \tilde{\delta}\right) &= (I+\rho M)^{-1}\left[y_1 - y_2\tilde{f}_1 - \underline{X}\tilde{\delta}\right] \\
&= (I+\rho M)^{-1}\left[(I+\rho_0 M)(u_1 - f_{1,0}u_2) - \Upsilon_1(\tilde{\vartheta} - \vartheta_0)\right],
\end{aligned}$$

and consequently

$$\begin{aligned} n^{-1}\epsilon^{+}(\rho,\tilde{f}_1,\tilde{\delta})' A\epsilon^{+}(\rho,\tilde{f}_1,\tilde{\delta}) &= n^{-1}\left(\tilde{\vartheta}-\vartheta_0\right)'\Upsilon_1'\left(I+\rho M\right)^{-1}A\left(I+\rho M\right)^{-1}\Upsilon_1\left(\tilde{\vartheta}-\vartheta_0\right) \\ &\quad -n^{-1}2\left(\tilde{\vartheta}-\vartheta_0\right)'\Upsilon_1'\left(I+\rho M\right)^{-1}A\mathcal{M}\left(\rho\right)\left(u_1-f_{0,1}u_2\right) \\ &\quad +n^{-1}\left(u_1-f_{1,0}u_2\right)'\mathcal{M}\left(\rho\right)A\mathcal{M}\left(\rho\right)\left(u_1-f_{1,0}u_2\right) \\ &= I+II+III. \end{aligned}$$

By Assumption 8 it follows that $\sup_{\rho}\left|n^{-1}\Upsilon_1'(I+\rho M)^{-1}A(I+\rho M)^{-1}\Upsilon_1-\mathscr{U}_{\Upsilon_1\Upsilon_1}(\rho))\right|\to 0$ a.s. for some uniformly bounded $U_{\Upsilon_1\Upsilon_1}(\rho)$. Thus I converges to zero uniformly in probability. Uniform convergence in probability to zero for the second term, II, follows by noting that $\Upsilon_1'A\left(I+\rho_0 M\right)\mathcal{M}\left(\rho\right)$ is $\mathcal{Z}_n^*$ measurable and uniformly bounded by, e.g., Remark A.1 in Kelejian and Prucha (2004) t such that pointwise convergence follows from Assumption 8. Uniform convergence follows from Lipschitz continuity as before.

Finally, using Lemmas B.2 and Assumption 8 it follows that

$$\sup_{\rho}\left|n^{-1}\left(u_1-f_{0,1}u_2\right)'\mathcal{M}\left(\rho\right)A\mathcal{M}\left(\rho\right)\left(u_1-f_{0,1}u_2\right)-\operatorname{tr}\left(\mathcal{M}\left(\rho\right)A\mathcal{M}\left(\rho\right)\Omega_0\right)\right|\to 0 \text{ a.s.}$$

Then the consistency of $\tilde{\rho}$ follows from Lemma 3.1 in Poetscher and Prucha (1997), and analogous arguments as those used for part (i) of the theorem. □

Proof of Theorem 4.2. First show consistency of $\hat{\gamma}$. Define $\Upsilon_1=[X,y_2]$ and $\vartheta=(\delta,f_1)$. observing from (8) that $y_1-f_{0,1}y_2-\underline{X}\delta_0=(I+\rho_0 M)(u_1-f_{0,1}u_2)$ we have

$$\begin{aligned} \tilde{u} &= \left(I_n+\tilde{\rho}M\right)^{-1}\left(y_1-\Upsilon_1\tilde{\vartheta}\right) \\ &= \left(u_1-u_2f_{1,0}\right)+\left(\mathcal{M}\left(\tilde{\rho}\right)-I\right)\left(u_1-u_2f_{1,0}\right) \\ &\quad +\mathcal{M}\left(\tilde{\rho}\right)\Upsilon_1\left(\vartheta_0-\tilde{\vartheta}\right) \end{aligned}$$

such that

$$\begin{aligned}
\sum_{c=1}^{C}\tilde{u}_c'\tilde{u}_c 1\left\{\tau_c=j\right\} &= \sum_{c=1}^{C}\left(u_{c1}-u_{c2}f_{0,1}\right)'\left(u_{c1}-u_{c2}f_{0,1}\right)1\left\{\tau_c=j\right\}\\
&\quad+\sum_{c=1}^{C}\left(u_{c1}-u_{c2}f_{0,1}\right)'\left(\mathcal{M}_c\left(\tilde{\rho}\right)-I\right)^2\left(u_{c1}-u_{c2}f_{0,1}\right)1\left\{\tau_c=j\right\}\\
&\quad+\sum_{c=1}^{C}\left(\vartheta_0-\tilde{\vartheta}\right)'\Upsilon_1'\mathcal{M}\left(\tilde{\rho}\right)'\mathcal{M}\left(\tilde{\rho}\right)\Upsilon_1\left(\vartheta_0-\tilde{\vartheta}\right)1\left\{\tau_c=j\right\}\\
&\quad+2\sum_{c=1}^{C}\left(u_{c1}-u_{c2}f_{0,1}\right)'\left(\mathcal{M}_c\left(\tilde{\rho}\right)-I\right)\left(u_{c1}-u_{c2}f_{0,1}\right)1\left\{\tau_c=j\right\}\\
&\quad+2\sum_{c=1}^{C}\left(u_{c1}-u_{c2}f_{0,1}\right)'\left(\mathcal{M}_c\left(\tilde{\rho}\right)-I\right)'\mathcal{M}_c\left(\tilde{\rho}\right)\Upsilon_1\left(\vartheta_0-\tilde{\vartheta}\right)1\left\{\tau_c=j\right\}\\
&\quad+2\sum_{c=1}^{C}\left(u_{c1}-u_{c2}f_{0,1}\right)'\mathcal{M}_c\left(\tilde{\rho}\right)\Upsilon_1\left(\vartheta_0-\tilde{\vartheta}\right)1\left\{\tau_c=j\right\}\\
&= I+II+III+IV+V+VI.
\end{aligned}$$

Now consider the components of $(N_j-q-1)^{-1}\sum_{c=1}^{C}\tilde{u}_c'\tilde{u}_c 1\left\{\tau_c=j\right\}$. For I, note that

$$\begin{aligned}
n^{-1}E\left[I\right] &= n^{-1}\sum_{c=1}^{C}E\left[E\left[\left(u_{c1}-u_{c2}f_{0,1}\right)'\left(u_{c1}-u_{c2}f_{0,1}\right)|\mathcal{Z}_n^*\right]1\left\{\tau_c=j\right\}\right]=\\
&= \sigma_0^2\left(1+f_{1,0}^2\right)n^{-1}\sum_{c=1}^{C}n_c\rho_{0,\tau_c}^2 1\left\{\tau_c=j\right\}\\
&= \sigma_0^2\left(1+f_{1,0}^2\right)\rho_{j,0}^2 n^{-1}\sum_{c=1}^{C}n_c\\
&= \sigma_0^2\left(1+f_{1,0}^2\right)\rho_{j,0}^2(N_j/n)\to\sigma_0^2\left(1+f_{1,0}^2\right)\rho_{j,0}^2 w_j
\end{aligned}$$

such that $\left(N_j-\underline{q}-1\right)^{-1}E\left[I\right]=[n/(N_j-\underline{q}-1]n^{-1}E\left[I\right]\to\sigma_0^2\rho_{j,0}^2\left(1+f_{1,0}^2\right)$ since $n/\left(N_j-q-1\right)\to 1/\omega_j$. By Assumption 8 it follows that $I\to\sigma_0^2\rho_{j,0}^2\left(1+f_{1,0}^2\right)$ a.s. uniformly in the parameter space for $j=1,\ldots,J-1$, and $I\to\sigma_0^2\left(1+f_{1,0}^2\right)$ a.s. for $j=J$ For II, IV and V consider

$$\begin{aligned}
\left(\mathcal{M}\left(\tilde{\rho}\right)-I\right) &= \left(I+\tilde{\rho}M\right)^{-1}M\left(\rho_0-\tilde{\rho}\right)\\
&= \operatorname{diag}_c\left(-\frac{1}{\left(n_c-1-\tilde{\rho}\right)}I_c^*+\frac{1}{1+\tilde{\rho}}J_c^*\right)\left(\rho_0-\tilde{\rho}\right)
\end{aligned}$$

such that for II

$$\left(\mathcal{M}_c\left(\tilde{\rho}\right)-I\right)^2=\operatorname{diag}_c\left(\frac{1}{\left(n_c-1-\tilde{\rho}\right)^2}I_c^*+\frac{1}{\left(1+\tilde{\rho}\right)^2}J_c^*\right)\left(\rho_0-\tilde{\rho}\right)^2.$$

Then using the notation

$$C(\rho) = \operatorname{diag}_c \left(\frac{1\{t_c = j\}}{(n_c - 1 - \rho)^2} I_c^* + \frac{1\{t_c = j\}}{(1+\rho)^2} J_c^* \right)$$

in line with Assumption 8 shows that

$$\begin{aligned} n^{-1} II &= (\rho_0 - \tilde{\rho})^2 n^{-1} (u_1 - u_2 f_{1,0})' C(\tilde{\rho}) (u_1 - u_2 f_{0,1}) \\ &\leq (\rho_0 - \tilde{\rho})^2 \sup_\rho \left| n^{-1} (u_1 - u_2 f_{0,1})' C(\rho) (u_1 - u_2 f_{0,1}) - E\left[n^{-1} \operatorname{tr}(C(\rho)\Omega_0) \right] \right| \\ &\quad + (\rho_0 - \tilde{\rho})^2 \sup_\rho \left| \left[E\left[n^{-1} \operatorname{tr}(C(\rho)\Omega_0) \right] - \lim_n E\left[n^{-1} \operatorname{tr}(C(\rho)\Omega_0) \right] \right] \right| \\ &\quad + (\rho_0 - \tilde{\rho})^2 \sup_\rho \left| \lim_n E\left[\operatorname{tr}(C(\rho)\Omega_0) \right] \right| \\ &= o_p(1), \end{aligned}$$

observing that $\rho_0 - \tilde{\rho} = o_p(1)$ by Theorem 4.1(ii), that the first and second supremum converges to zero a.s. by Assumption 8(i) and that $\lim_n E[\operatorname{tr}(C(\rho)\Omega_0)]$ is uniformly bounded by Assumption 8(ii). Arguments analogous to these show that the remaining terms of $n^{-1}\sum_{c=1}^{C} \tilde{u}_c' \tilde{u}_c 1\{\tau_c = j\}$ converge to zero in probability. The details are omitted. This establishes that

$$\hat{\gamma}_j = \left(N_j - \underline{q} - 1\right)^{-1} \sum_{c=1}^{C} \tilde{u}_c' \tilde{u}_c 1\{\tau_c = j\} \rightarrow_p \sigma_0^2 \rho_{0,j}^2 \left(1 + f_{0,1}^2\right).$$

To obtain the limiting distribution note that $Q_n(\phi, \gamma)$ defined in (15) can be written as

$$Q_n(\phi, \gamma) = m_n(\phi, \gamma)' \Xi_n^{-1} m_n(\phi, \gamma)$$

where Ξ_n is defined in (14). Using a mean value expansion around ϕ_0 we obtain

$$0 = \sqrt{n} \frac{\partial Q_n\left(\hat{\phi}, \gamma\right)}{\partial \phi} = \frac{\partial m_n(\phi_0, \gamma)'}{\partial \phi} \Xi_n^{-1} \sqrt{n} m_n(\phi_0, \gamma) + \frac{\partial m_n(\phi, \gamma)'}{\partial \phi} \Xi_n^{-1} \frac{\partial m_n\left(\bar{\phi}, \gamma\right)}{\partial \phi} \sqrt{n}\left(\hat{\phi} - \phi_0\right)$$

where $\left\|\bar{\phi} - \phi_0\right\| \leq \left\|\hat{\phi} - \phi_0\right\|$ and with some abuse of notation it is understood that $\bar{\phi}$ differs among rows of $m_n(.)$. Using the explicit derivatives given in Section D it follows from Assumption 8(i) and (ii) that for any sequence γ_n with $\gamma_n \rightarrow_p \gamma_*$ that $\partial m_n\left(\bar{\phi}, \gamma_n\right)/\partial \phi \rightarrow_p G$ where G is fixed matrix and $\Xi_n \rightarrow_p \Xi$ with Ξ a square matrix and where G is full column rank and Ξ is full rank. This implies that

$$\sqrt{n}\left(\hat{\phi} - \phi_0\right) = \left(\left(G'\Xi^{-1}G\right)^{-1} G'\Xi^{-1} + o_p(1) \right) \sqrt{n} m_n(\phi_0, \gamma_n).$$

It remains to establish the limiting distribution of $\sqrt{n} m_n(\phi_0, \gamma_n)$ where γ_n is a possibly random sequence with $\gamma_n \rightarrow_p \gamma_*$. To apply the results of Kuersteiner and Prucha (2020) note that the moment vector $m_n(.)$ can be represented as the sum over the linear and quadratic terms by defining $h_{i,1} = (h_i', 0)'$ and $a_{ij,1} = (0, ..., 0, a_{ij})'$ with h_i the i-th row of H and a_{ij} the i,j-th element of A. Then, $m_n(\phi, \gamma_n) = \sum_{i=1}^n h_{i,1} u_{i,*1}^+ + \sum_{j,i=1}^n a_{ij,1} u_{i,*1}^+ u_{j,*1}^+$ has the same form as Equation (23) in Kuersteiner and Prucha (2020). It is then sufficient to check that the conditions for Proposition 3 in Kuersteiner and Prucha (2020) hold. Proposition 3 requires that Assumptions 1-3 of Kuersteiner and Prucha (2020) hold. We abbreviate assumptions in Kuersteiner and Prucha (2020) with A-KP to avoid confusion with assumptions in this paper.

The moment bounds of A-KP 1(i) hold by Assumption 2 and Lemma 2.1 which implies that u_{it} and μ_i have bounded moments, and Assumption 7(i) which bounds the moments of x_{it} and z_{it}. A-KP 1(ii), Eq 15, holds by Lemma B.3, A-KP 1(ii), Eq 16 holds by Assumption **??** and A-KP 1(ii), Eq 17 holds by Lemma B.3, and the parametric restrictions on the conditional cross-sectional variances hold by Assumption 3.

Now turning to A-KP 2, the σ-field that corresponds to $\mathcal{B}_{n,t}$ is $\mathcal{A}_{n,(t-1)n}$ defined in (39). Since $\mathcal{Z}_n^* \subset \mathcal{A}_{n,(t-1)n}$, the fact that h_i and a_{ij} are assumed measurable w.r.t to $\mathcal{Z}_n \vee \mathcal{C} \subset \mathcal{Z}_n^*$ in Assumption 7(i) and (iii), and uniformly bounded in Assumption 7(i), A-KP 2(i) and (ii) hold. We also normalize $f_2 = 1$ and restrict the parameter space of f_1 to a compact interval in Assumption 4 such that A-KP 2(iii) holds.

Finally, consider A-KP 3 which is identical to Assumption 8. In summary, all conditions of Proposition 3 in Kuersteiner and Prucha (2020) hold. It follows that $\sqrt{n} m_n(\phi_0, \gamma) \to^d V_\varrho^{1/2} \xi$ where $\xi \sim N(0, I_{q+2})$ For any conformable, matrix A it follows that $\sqrt{n} A m_n(\phi_0, \gamma) \to^d (A'VA)^{1/2} \xi$. By choosing $A = \left(G'\Xi^{-1}G\right)^{-1} G'\Xi^{-1}$ the claim of the theorem is proven. □

D Gradient Vector

In this section, we derive the gradient vector for the two-step-GMM estimator. Recall that $\phi = [\rho, f_1, \delta']'$,

$$
\begin{aligned}
u_1^+(\phi, \gamma) &= \Omega(\gamma)^{-1/2}(I + \rho M)^{-1}\left(y_1 - f_1 y_2 - \underline{X}\delta\right) \\
&= \Omega(\gamma)^{-1/2}(I + \rho M)^{-1}\left((f_{10} - f_1)\mu^* + \underline{\mathrm{X}}(\delta(f_1, \rho_0, \beta_0) - \delta) + (I + \rho_0 M)(u_1 - f_1 u_2)\right)
\end{aligned}
$$

and that $y_t = \mu^* f_{0,t} + \underline{X}\beta_{0,t} + (I + \rho_0 M)u_t$.

$$
\frac{\partial m_n(\phi, \gamma)/\sqrt{n}}{\partial \phi'} = \frac{1}{n}\begin{bmatrix} H' \frac{\partial u_1^+}{\partial \phi'} \\ 2u_1^{+\prime} A \frac{\partial u_1^{+\prime}}{\partial \phi'} \end{bmatrix} = \frac{1}{n}\begin{bmatrix} H' \frac{\partial u_1^+}{\partial \rho} & H' \frac{\partial u_1^+}{\partial f_1} & H' \frac{\partial u_1^+}{\partial \delta'} \\ 2u_1^{+\prime} A \frac{\partial u_1^{+\prime}}{\partial \rho} & 2u_1^{+\prime} A \frac{\partial u_1^{+\prime}}{\partial f_1} & 2u_1^{+\prime} A \frac{\partial u_1^+}{\partial \delta'} \end{bmatrix}.
$$

The derivative of $u^{+}(\phi,\gamma)$ w.r.t $\phi=(\rho,f_1,\delta)$ are

$$\partial\frac{u^{+}(\phi,\gamma)}{\partial f_1} = -\Omega(\gamma)^{-1/2}(I+\rho M)^{-1}y_2$$
$$= -\Omega(\gamma)^{-1/2}(I+\rho M)^{-1}[\mu^{*}+\underline{X}\beta_{0,2}+(I+\rho_0 M)u_2].$$

$$\partial\frac{u^{+}(\phi,\gamma)}{\partial\rho} = -\Omega(\gamma)^{-1/2}(I+\rho M)^{-1}M(I+\rho M)^{-1}(y_1-f_1y_2-\underline{X}\delta)$$
$$= -\Omega(\gamma)^{-1/2}(I+\rho M)^{-1}M(I+\rho M)^{-1}$$
$$\times((f_{10}-f_1)\mu^{*}+\underline{X}(\delta(f_1,\rho_0,\beta_0)-\delta)+(I+\rho_0 M)(u_1-f_1u_2))$$

$$\partial\frac{u^{+}(\phi,\gamma)}{\partial\delta} = -\Omega(\gamma)^{-1/2}(I+\rho M)^{-1}\underline{X}.$$

It thus is obvious that $u^{+}(\phi,\gamma)$ and each element of $\frac{u_1^{+}(\phi,\gamma)}{\partial\phi}$ can be written as linear combinations of $S(\phi,\gamma)\Upsilon$, $S(\phi,\gamma)u_t$, where $S(\phi,\gamma)$ is an element or the product of elements of the matrix set in (38), and $\Upsilon\in\{H,\mu^{*},\underline{X}\}$. Applying Corollary B.1 and Assumption 8 , the gradient vector converges uniformly to its expected value.