Identification and Estimation of Group-Level Partial Effects*

Kenichi Nagasawa†

December 17, 2018

Job Market Paper

Latest version here.

Abstract

This paper presents a new identification result for causal effects of group-level variables when agents select into groups. The model allows for group selection to be based on individual unobserved heterogeneity. This feature leads to correlation between group-level covariates and unobserved individual heterogeneity. Whereas many of the existing identification strategies rely on instrumental variables for group selection, I introduce alternative identifying conditions which involve individual-level covariates that "shift" the distribution of unobserved heterogeneity. I use these conditions to construct a valid control function. The key identifying requirements on the observable "shifter" variables are likely to hold in settings where a rich array of individual characteristics are observed. The identification strategy is constructive and leads to a semiparametric, regression-based estimator of group-level causal effects, which I show to be consistent and asymptotically normal. A simulation study indicates good finite-sample properties of this estimator. I use my results to re-analyze the effects of school/neighborhood characteristics on student outcomes, following the work of Altonji and Mansfield (2018), and I find that moving from a 10th to 90th percentile school/neighborhood increases wages by 17.58%.

*I would like to express my deep appreciation to Matias Cattaneo for his continual advice and encouragement. I would also like to thank Lutz Kilian, Andreas Hagemann, and Rocío Titiunik for valuable feedback and encouragement. I also benefited from helpful discussions with Sebastian Calonico, Max Farrell, Yingjie Feng, Michael Jansson, Toshiaki Komatsu, Xinwei Ma, Dhiren Patki, and Gonzalo Vazquez-Bare. I am thankful to Joseph Altonji and Richard Mansfield for sharing their dataset and code.

†The University of Michigan, Department of Economics.

1 Introduction

Policy makers often design interventions to influence individual outcomes through group-level variables. For instance, a government may relocate disadvantaged children to higher quality schools to improve their academic performance. Given their potential impact, many studies in economics have sought to evaluate group-level policy interventions (see Durlauf, 2004; Durlauf and Ioannides, 2010; Graham, 2018, and references therein). Nevertheless, estimation of group-level treatment effects is challenging. The problem is that individuals select into groups in part based on their unobserved characteristics, and this sorting causes systematic dependence among group-level variables and those individual characteristics. Therefore, comparing outcomes across groups without accounting for differences in unobserved heterogeneity is subject to selection bias.

This paper presents a novel identification result for group-level partial effects using observable variables that "shift" the distribution of unobserved individual heterogeneity. Informally, a random vector W is a "shifter" for another (unobservable) random vector Θ if the conditional distribution of Θ, given $W = w$, varies sufficiently with w. A key insight of this paper is that controlling for group-level distributions of observable shifters accounts for variation in unobserved individual heterogeneity. Thus, if there remains enough variation in group-level variables conditional on the control variables, group-level causal effects become identifiable.

For simplicity, suppose that unobserved individual heterogeneity Θ takes two values, "high" and "low," denoted by θ_H and θ_L. Then, an observable W is a shifter for Θ if $\Pr(\Theta = \theta_H|W = w)$ is a non-constant function of w, or equivalently, if there exist w_1, w_2 such that the 2×2 matrix $\mathbf{\Pi} = [\Pr(\Theta = \theta_l|W = w_k)]_{k=1,2;l=H,L}$ is non-singular. The availability of a shifter is useful because a group-level distribution of W can be written as a linear transformation of the distribution of Θ within the same group, where the matrix of this linear transformation is $\mathbf{\Pi}$. Since $\mathbf{\Pi}$ is non-singular, there exists a one-to-one mapping between group-level distributions of the unobserved variable Θ and the observable shifter W. Therefore, we can use an observable shifter to control for across-group variation in unobserved heterogeneity distributions.

To fix ideas, consider a setting where students choose schools. Academically motivated students may prefer high quality schools, and when they sort into schools based on this preference, there will be a positive correlation between school-level teacher quality and the proportion of highly moti-

vated students within the school. Here, student's motivation corresponds to unobserved individual heterogeneity and school-level teacher quality is the group-level variable of interest. A possible shifter for academic motivation is educational attainment of a child's mother. It seems reasonable to assume that the higher is a mother's educational attainment, the more motivated her child tends to be.

Now suppose that student's motivation takes two values, high and low. Then, mother's education serves as a shifter if a child of a college graduate mother has a larger probability of being highly motivated than a student whose mother has only a high school diploma. Provided that mother's education is a valid shifter, there exists a one-to-one mapping between a school-level fraction of college graduate mothers and the distribution of student types within the school, which enables us to control for across-school variation in student's motivation. If school-wide teacher quality has independent variation from the school-level distribution of mother's education, then ceteris paribus effects of teacher quality are identifiable. Note that mother's education is not an instrumental variable (IV): a valid IV would be independent of student's motivation and be related to teacher quality.

To provide further intuition on the above argument, I describe it with simple equations. Denote an outcome of interest by Y_{is} (e.g., test score), where i and s index agent and group, respectively, group-level covariates of interest by X_s (e.g., school-wide teacher quality), an observable shifter by W_i (e.g., mother's education), and unobserved individual heterogeneity by Θ_i (e.g., academic motivation). Like the above discussion, assume Θ_i takes two values, θ_H and θ_L (e.g., high and low motivation). Also, let J_i be individual i's group choice. Here, the interest lies in the causal effect of teacher quality on test scores. The regression equation turns out to be

$$Y_{iJ_i} = \alpha + X'_{J_i}\beta + W'_i\gamma + \delta \Pr(\Theta_i = \theta_H | X_{J_i}, J_i) + \varepsilon_{iJ_i}, \qquad \mathbb{E}[\varepsilon_{is} | X_s, W_i, J_i = s] = 0 \tag{1}$$

where subscript J_i is present because a researcher only observes the outcome for the group individuals selected in the data. Since $\Pr(\Theta_i = \theta_H | X_s, J_i = s)$ cannot be estimated, its dependence on X_s hinders the identification of β. In the school example, $\Pr(\Theta_i = \theta_H | X_s, J_i = s)$ corresponds to the fraction of highly motivated students in a school with teacher quality X_s. In this framework, the inability to control for a fraction of high motivation types confounds naive estimates of

teacher-quality effects.

As noted above, the observable shifter condition implies that there exists a linear map from the distribution of the observable shifter to the unobserved heterogeneity distribution, i.e., $\Pr(\Theta_i = \theta_H | X_s, J_i = s) = \sum_{l=1}^{2} d_l \Pr(W_i = w_l | X_s, J_i = s)$ for some $d_1, d_2 \in \mathbb{R}$. Then, (1) becomes

$$Y_{iJ_i} = \alpha + X'_{J_i}\beta + W'_i\gamma + \sum_{l=1}^{2} \delta d_l \Pr(W_i = w_l | X_{J_i}, J_i) + \varepsilon_{iJ_i}$$

where the error term ε_{iJ_i} is uncorrelated with all the other right-hand side variables. Since $[\Pr(W_i = w_l | X_s, J_i = s)]_{l=1,2}$ is identifiable from the data (e.g., school-level fractions of college graduate and high school graduate mothers), β becomes identifiable.

I generalize the argument based on the two-point supported distribution of Θ to one with a general distribution. To formalize the idea of shifter, I use the notion of statistical completeness, which has been applied in a wide range of nonparametric identification problems. I argue that the availability of a shifter is a reasonable assumption, especially when a researcher observes a rich array of individual-level characteristics.

When completeness is used for identification, estimation often faces ill-posed inverse problems, which may lead to poor finite-sample properties of estimators. The estimator in this paper circumvents this problem by applying the completeness assumption to estimation of a nuisance parameter of the model. By additive separability of the nuisance function from the parameter of interest, the estimator for group-level partial effects converges at a rate of the square root of the sample size as happens in partially linear regression models.

For implementation, a researcher first picks a set of basis functions and computes the group-means of the observable shifters transformed by those basis functions. She then runs a linear regression of the outcome on the group-level covariates, individual-level regressors, and the group-means computed in the first step. I show that this procedure leads to a consistent and asymptotically normal estimator, and I provide a consistent variance estimator. Monte Carlo experiments demonstrate that the proposed estimator has low mean squared errors and that the associated confidence intervals have high coverage accuracy in samples of moderate size.

I apply this method to the problem of studying the effects of school/neighborhood characteristics on years of post-secondary education and adulthood wages. I build on the analysis of Altonji and

Mansfield (2018), who also use a control function approach. This paper shows that the estimation equation arising from my identification result encompasses that of Altonji and Mansfield. Thus, my estimator is more robust to possible mis-specifications. I find that the two estimators produce very similar estimates, which supports the results in Altonji and Mansfield.

This paper employs assumptions distinct from those of existing identification strategies. In particular, I do not require natural/quasi experimental variation as often used in IV methods. Instead, I use the observable shifter condition, which can hold in many applications of interest. Thus, I provide an alternative approach to identifying group-level treatment effects. In addition, my identification result has wide applicability since the argument based on the shifter condition extends to non-linear and nonseparable models in a straightforward manner. From a theoretical perspective, this paper achieves new identification results using unexploited yet empirically relevant features in triangular models, such as group structures. Lastly, this paper applies statistical completeness in a novel way to develop a control function approach. To elaborate on these contributions, I now review the related literature.

1.1 Related Literature

This paper contributes to the large empirical literature that examines causal effects of neighborhood, school, and, more generally, group-level characteristics.[1] A non-trivial challenge in this literature is how to control for endogeneity arising from agents selecting into groups (Durlauf, 2004; Graham, 2018). Many studies exploit exogenous variation in group selection via instrumental variables (e.g., Angrist, Pathak, and Walters, 2013; Kling, Liebman, and Katz, 2007; Ludwig, Duncan, and Hirschfield, 2001; Oreopoulos, 2003), some studies use detailed data on individual choice behavior to control for selection bias (e.g., Abdulkadiroğlu, Pathak, Schellenberg, and Walters, 2017; Dale and Krueger, 2002), and other studies use aggregation to mitigate influence of selection (e.g., Card and Rothstein, 2007). In contrast, I take a control function approach that complements the existing methods by exploiting a novel, yet empirically relevant identifying condition.

This paper is most closely related to recent work by Altonji and Mansfield (2018), who also

[1] A partial list includes Aaronson (1998); Abdulkadiroğlu, Pathak, and Walters (2018); Altonji, Elder, and Taber (2005); Angrist and Lang (2004); Chetty and Hendren (2018); Chetty, Hendren, and Katz (2016); Dobbie and Fryer (2011); Gould, Lavy, and Paserman (2004); Hanushek, Kain, and Rivkin (2009); Hoxby (2000). Also, Chetverikov, Larsen, and Palmer (2016) study effects of group-level endogenous variables. They do not consider endogeneity from selection and focus on other issues.

employ a control function method. They show that group-means of individual covariates play the role of control functions in their model. In contrast, I use group-means of *transformations* of individual covariates as control functions, where these transformations are basis functions (e.g., polynomials and splines). Whereas Altonji and Mansfield exploit specific functional forms in their model to justify their control function method, I use restrictions on the conditional distribution of individual-level unobserved heterogeneity given observables to construct a control function. An advantage of my approach is that the identification argument extends to non-linear, nonseparable models in a relatively straightforward manner and thus my results apply to a wide class of models. In addition, the estimation equation arising from my identification result encompasses that of Altonji and Mansfield. In fact, when certain regression functions in the model are linear, the two approaches produce effectively the same identification result.

Another strand of the related literature is the one on control function methods, which is also connected to triangular models. A general triangular system takes the form

$$Y = m(X, \varepsilon)$$
$$X = h(Z, \eta)$$

where Y is the outcome of interest, X is potentially endogenous, and Z is an instrument independent of the unobserved heterogeneity (ε, η). In his seminal work, Heckman (1974, 1979) develops a control function approach where X is a binary variable, using additive separability and joint normality of (ε, η). Subsequent papers (e.g., Dahl, 2002; Das, Newey, and Vella, 2003; Dubin and McFadden, 1984; Lee, 1983) extend the model in different directions. Examples include allowing for multinomial X and weakening parametric distributional assumptions. Also, Newey, Powell, and Vella (1999) develop a control function method for nonparametric triangular systems and Blundell and Powell (2004) use a control function to identify average partial effects in semiparametric limited dependent variable models. Recently, Chesher (2003) and Imbens and Newey (2009) exploit strict monotonicity of functions in unobserved heterogeneity to construct a control variate to identify the ceteris paribus effect of X on Y (see also Matzkin, 2016, and references therein). I consider a version of the triangular model that exploits availability of multiple measurements of groups as well as an observable shifter to construct a control function. In contrast, the earlier literature seems

to have largely focused on monotonicity restrictions and distributional assumptions on unobserved heterogeneity as identifying conditions. The use of empirically relevant features such as multiple measurements of groups offers a new approach to achieving identification in this class of models.

In addition, this paper is related to the growing literature on nonparametric identification using statistical completeness. Since the seminal work of Newey and Powell (2003), completeness has been applied to a wide range of econometric identification problems. Examples include nonparametric IV (e.g., Chernozhukov and Hansen, 2005; Darolles, Fan, Florens, and Renault, 2011; Hall and Horowitz, 2005; Newey and Powell, 2003), errors-in-variables models (Hu and Schennach, 2008), nonparametric discrete choice models with unobserved product characteristics (Berry and Haile, 2014), and nonseparable (dynamic) panel data models (e.g., Arellano, Blundell, and Bonhomme, 2017; Cunha, Heckman, and Schennach, 2010; Freyberger, 2018; Sasaki, 2015). See also the survey article Hu (2017) and references therein. My paper applies the completeness assumption in a novel way to construct a control function.

This paper shares with the literature on network formation the feature that agents form groups endogenously within models (e.g., Blume, Brock, Durlauf, and Jayaraman, 2015; de Paula, 2017; Graham, 2017, and see references therein). However, I consider a different group selection mechanism from those in the network literature. In particular, in my analysis, the utility function for group selection does not depend on other individuals' group choices and their characteristics. Therefore, my model excludes certain selection patterns that are possible under network formation models. Nonetheless, the model still covers many empirical settings of interest.

The remainder of the paper is organized as follows. Section 2 describes the econometric model and provides a heuristic discussion of the identification strategy. In Section 3, I formalize the identification idea. In Section 4, I propose a simple estimator of group-level partial effects and study its asymptotic properties. Section 5 discusses the empirical application and Monte Carlo experiments to examine finite-sample accuracy of the proposed estimator. Section 6 introduces two extensions of the model described in Section 2. Section 7 concludes. Details of the proofs can be found in the appendix.

2 Setup and Overview of Results

In this section, I describe the econometric model. A researcher observes "cities," indexed by $g = 1, \ldots, G$. Within each city, there exist groups and agents, indexed by $s \in \mathcal{S} \equiv \{1, \ldots, S\}$ and $i = 1, 2, \ldots, N$, respectively. The outcome of interest, denoted by Y_{isg}, takes the form

$$Y_{isg} = \beta' X_{sg} + \gamma' W_{ig} + \chi_{sg} + \omega_{ig} + \epsilon_{isg}, \qquad \mathbb{E}[\epsilon_{isg} | X_{sg}, W_{ig}, \chi_{sg}, \omega_{ig}] = 0 \tag{2}$$

where $\{X_{sg} : s \in \mathcal{S}\}$ and W_{ig} are observable variables for groups and individuals, respectively, $\{\chi_{sg} : s \in \mathcal{S}\}$ and ω_{ig} denote unobservable characteristics for groups and individuals, respectively, and $\{\epsilon_{isg} : s \in \mathcal{S}\}$ represents unobserved idiosyncratic terms. The focus of this paper is on group-level causal effects, captured by β in (2). To identify this parameter, I require variation in group-level variables. The presence of cities provides multiple measurements, or independent copies, of groups and agents, which enables identifying the distribution of group-level covariates and within-group distributions of individual variables. Below, I omit city index g to reduce notational burden when appropriate.

To ground the discussion on concrete terms, consider the following examples.

Example 1 (Residential Segregation and Youth Outcomes)**.** Graham (2018) surveys issues of residential segregation and its consequence on youth outcomes. In one example, the outcome Y_{is} measures adulthood wage, W_i denotes the indicator of a resident being a minority, and X_s represents the proportion of minorities in one's neighborhood. Here a neighborhood corresponds to a group. Unobservables that enter into the equation include resident's innate cognitive ability (ω_i) and distance from the city center to one's neighborhood (χ_s). The parameter of interest is β, which measures how the proportion of minorities in one's neighborhood affects an outcome of interest. □

Example 2 (Effects of School District/Neighborhood on Student Performance)**.** Altonji and Mansfield (2018) examine effects of school district/neighborhood on various student outcomes. In one of their empirical specifications, Y_{is} is years of post-secondary education, X_s includes teacher-student ratio and distance to four-year college, W_i contains student's scores on standardized tests and parents' years of education, ω_i includes how much parents value child's academic learning, and χ_s

represents unobserved characteristics of school. Here, a school district/neighborhood represents a group. □

Example 3 (Effects of Hospital Ownership on Quality of Care)**.** Sloan, Picone, Taylor, and Chou (2001) study whether private, for-profit hospitals have lower quality of care. Since different types of patients may select into for-profit hospitals compared to public ones, selection bias is of potential concern. Here, a hospital corresponds to a group, the outcome Y_{is} is patient's condition after hospitalization, X_s includes the indicator of whether the hospital is private and for-profit, and W_i includes measures of health conditions before hospitalization. □

In these examples, a researcher observes the outcome variable only for the group an individual selects into. To be precise, let $J_{ig} \in \mathcal{S}$ be the group membership of agent i. Then, we only observe $Y_{ig} := \sum_{s \in \mathcal{S}} Y_{isg} \mathbb{1}\{J_{ig} = s\}$. I model the group membership determination as following.

$$J_{ig} = J(\Theta_{ig}, A_{1g}, \ldots, A_{Sg}, \eta_{i1g}, \ldots, \eta_{iSg}) \tag{3}$$

where $J(\cdot)$ is an unknown function belonging to a nonparametric class, Θ_{ig} is individual heterogeneity affecting the selection, $\{A_{sg} : s \in \mathcal{S}\}$ are characteristics of groups affecting agents' group choice, and $\{\eta_{isg} : s \in \mathcal{S}\}$ are idiosyncratic terms. One example of the selection equation $J(\cdot)$ that fits into this framework is the widely used random utility discrete choice model. In that setting, $J(\Theta_i, A_1, \ldots, A_S, \eta_{i1}, \ldots, \eta_{iS}) = \arg\max_{s \in \mathcal{S}} V_i(s)$ and $V_i(s) = \Theta_i' A_s + \eta_{is}$, where $V_i(s)$ represents agent i's utility choosing the group s, and the utility is specified as the group-level features A_s weighted by the taste coefficient Θ_i plus the idiosyncratic term η_{is}. Random coefficient specification allows for heterogeneous tastes for group characteristics and accommodates complicated selection patterns.

A key restriction on the selection function $J(\cdot)$ is that, loosely speaking, if a researcher were to observe all the variables, she could compare groups with different X_s holding within-group distributions of Θ_i constant. That is, there must exist separate variation in X_s from within-group distributions of individual characteristics. This condition holds, for instance, in the random utility discrete choice model satisfying the following: given $x_1, x_2 \in \text{supp}(X_s)$, there exist $\mathbf{a}^k \in \text{supp}(\{A_s : s \in \mathcal{S}\} | X_s = x_k), k = 1, 2$ satisfying $a_s^1 - a_{s'}^1 = a_s^2 - a_{s'}^2$ for all pairs of $(s, s') \in \mathcal{S}^2$. Alternative conditions are possible, and I formalize this assumption as non-singularity of some conditional

variance in Section 3. Also, I accommodate the possibility that any or all of $(\Theta_i, \{\eta_{is}, A_s : s \in \mathcal{S}\})$ in the selection equation is not observed, and instead I impose that the observable W_i has non-trivial relationship with the preference heterogeneity Θ_i.

An important feature of the selection equation (3) is that it does not depend on choices and characteristics of other individuals, which excludes the types of models considered in network formation literature. Yet, this does not exclude equilibrium effects. For instance, X_s can include features of the distribution of individual characteristics within group s, which can be empirically important since the coefficients on such variables may represent "peer effects," e.g., the effect of school-level minority fraction on student performance. As I will outline below, the control function method I propose uses group-level averages of (transformed) variables W_i. If X_s includes some features of the within-group distribution of a subvector of W_i (e.g., mean), the subvector needs to be excluded from the construction of the control functions.

Given the outcome and selection equations, I now state the following sampling assumptions to complete the model description.[2]

$$(\mathbf{X}_g, \boldsymbol{\chi}_g, \mathbf{A}_g) \overset{iid}{\sim} F, \qquad (W_{ig}, \omega_{ig}, \Theta_{ig}, \boldsymbol{\epsilon}_{ig}, \boldsymbol{\eta}_{ig}) \overset{iid}{\sim} H \tag{4}$$

$$(W_{ig}, \omega_{ig}, \Theta_{ig}, \boldsymbol{\epsilon}_{ig}, \boldsymbol{\eta}_{ig}) \perp\!\!\!\perp (\mathbf{X}_g, \boldsymbol{\chi}_g, \mathbf{A}_g) \tag{5}$$

$$(W_{ig}, \omega_{ig}, \boldsymbol{\epsilon}_{ig}) \perp\!\!\!\perp \boldsymbol{\eta}_{ig} | \Theta_{ig} \tag{6}$$

where $\mathbf{X}_g = \{X_{sg} : s \in \mathcal{S}\}$, $\boldsymbol{\chi}_g = \{\chi_{sg} : s \in \mathcal{S}\}$, $\mathbf{A}_g = \{A_{sg} : s \in \mathcal{S}\}$, $\boldsymbol{\epsilon}_{ig} = \{\epsilon_{isg} : s \in \mathcal{S}\}$, $\boldsymbol{\eta}_{ig} = \{\eta_{isg} : s \in \mathcal{S}\}$, and F and H are distribution functions. Restriction (4) states that the group-level features are i.i.d. copies across cities and individual-level variables are a random draw across individuals and cities.

Condition (5) requires independence between individual variables and group-level variables. One way to rationalize this independence assumption is that "nature" first draws group-level characteristics, then subsequently samples individual characteristics whose distributions are independent of group-level variables, and finally individuals select into groups. In Example 1, this scenario means that first neighborhoods are formed, then individuals are drawn from the distribution that does not depend on neighborhood characteristics, and after the realizations of group- and individual-

[2]These conditions are slightly stronger than those in Section 3. I maintain them here for ease of exposition.

level variables, agents make decisions on which neighborhood to live in. This assumption seems reasonable under static frameworks.

Condition (6) states that the idiosyncratic term in the choice equation is independent of other individual-level variables conditional on the taste variable Θ_i. Independence is a strong assumption, but such restriction has been commonly used in the literature of discrete choice models (e.g., Briesch, Chintagunta, and Matzkin, 2010). Also, here the independence only needs to hold conditioning on the taste heterogeneity.

2.1 Identification Problem

In this subsection, I describe a challenge in identifying the group-level effect β. Specifically, there are two sources of endogeneity, *omitted variable bias* and *selection bias*. The former occurs from unobservability of group-level characteristics χ_s and potential correlation between X_s and χ_s. Although this issue is practically relevant, it is well understood in the literature and this paper has little new insight to offer on this problem. Instead, I focus on selection bias and assume that there is no correlation between X_s and χ_s except in Section 6.1. In particular, I impose $\mathbb{E}[\chi_s|W_i, X_s, J_i = s] = 0$ for all s in this section.

For selection bias, the problem arises because the distribution of ω_i, the unobservable in the outcome equation (2), varies across groups and this variation is systematically related to X_s. In the school example, the unobservable ω_i represents student's motivation, and school characteristics affecting selection A_s contain the school-level teacher quality variable X_s. If highly motivated students prefer high quality educational programs, students with high ω_i are then more likely to be in schools with higher X_s. This sorting pattern will cause bias on the coefficient on X_s.

I now formalize selection bias in the model. Recall that we only observe the outcome for one group, i.e., $Y_i = \sum_{s \in \mathcal{S}} \mathbb{1}\{J_i = s\} Y_{is}$, and computing the regression function given observable variables yields

$$\mathbb{E}[Y_i|J_i = s, X_s, W_i] = \mathbb{E}[Y_{is}|J_i = s, X_s, W_i] = \beta' X_s + \gamma' W_i + \mathbb{E}[\omega_i|J_i = s, X_s, W_i].$$

Identifiability of β depends on whether the conditional expectation $\mathbb{E}[\omega_i|J_i = s, X_s, W_i]$ is constant in X_s. This model implication is closely related to the characterization of selection bias in Heckman

(1976). He formulates that $\mathbb{E}[Y|Z, \{\text{selection rule}\}] = Z'\beta + \mathbb{E}[\epsilon|Z, \{\text{selection rule}\}]$, where Z is a vector of covariates and ϵ is an error term. In my model $\{J_i = s\}$ corresponds to the "selection rule." Heckman points out that many econometric models have this characterization and that agents' self-selection prevents identification of β when the term $\mathbb{E}[\epsilon|Z, \{\text{selection rule}\}]$ is non-constant.

To gain some intuition on how the selection rule operates in my model, take $A_s = X_s$ and $J_i = \arg\max_{s \in \mathcal{S}} \{\Theta_i' X_s\}$ (i.e., I set $\eta_{is} = 0$ for all i and s). Then, conditioning on $\{J_i = s\}$ and $\mathbf{X}$ implies $\Theta_i \in \mathcal{R}_s(\mathbf{X})$ where $\mathcal{R}_s(\mathbf{X}) = \{\Theta : (X_s - X_{s'})'\Theta > 0 \text{ for all } s' \neq s\}$ is a region formed by intersections of half planes. See Figure 1 for a simple illustration. Then, for the conditional expectation $\mathbb{E}[\omega_i|J_i = s, X_s, W_i]$,

$$\begin{aligned}\mathbb{E}[\omega_i|J_i = s, X_s, W_i] &= \mathbb{E}[\mathbb{E}[\omega_i|J_i = s, \mathbf{X}, W_i]|J_i = s, X_s, W_i] \\ &= \mathbb{E}[\mathbb{E}[\omega_i|\Theta_i \in \mathcal{R}_s(\mathbf{X}), W_i]|J_i = s, X_s, W_i] \end{aligned} \tag{7}$$

where the second equality follows from $(\omega_i, \Theta_i, W_i) \perp\!\!\!\perp \mathbf{X}$. Now, to see implications of the last display, suppose $\mathbb{E}[\omega_i|\Theta_i, W_i] = \mathbb{E}[\omega_i|W_i]$. Then (7) implies $\mathbb{E}[\omega_i|J_i = s, X_s, W_i] = \mathbb{E}[\omega_i|W_i]$ and therefore no selection bias occurs. This result resembles the one in Heckman (1976): if unobserved terms in outcome and selection equations (ω_i and Θ_i in this model) are independent, there is no endogeneity due to selection. On the other hand, if the conditional mean of ω_i given $\Theta_i = \theta$ and W_i non-trivially changes with θ, then the term $\mathbb{E}[\omega_i|J_i = s, X_s, W_i]$ varies with X_s since the shape of $\mathcal{R}_s(\mathbf{X})$ changes with X_s. The dependence of $\mathbb{E}[\omega_i|J_i = s, X_s, W_i]$ on X_s results in non-identification of β.

Note that the above argument continues to hold even if we deviate from the simplifying assumption $A_s = X_s$, provided that A_s is related to X_s, which is likely the case. For example, X_s measures school-level teacher quality, and students are likely to use it or its proxy as a basis of their school choice decision. Also, the conditional mean independence of ω_i given Θ_i is likely to fail in applications if ω_i contains unobserved heterogeneity that affects people's preference for group characteristics Θ_i, e.g., if student's motivation affects preference for school quality.

Example 1 (Continued)**.** Graham (2018) discusses potential sources of endogeneity that hinder identification of causal effects of neighborhood segregation of minorities. Among the conditions he discusses, the *no sorting on (individual) unobservables* condition is relevant to my paper. Roughly

speaking, this condition requires that individuals with the same observed covariates are similar across neighborhoods. It excludes, for example, racial minorities living in one neighborhood differ substantially in their cognitive skills from minorities in another neighborhood.

To formalize the "no sorting" condition, redefine γ and ω_i in (2) by projecting ω_i onto W_i to have $\text{Cov}(W_i, \omega_i) = 0$. This is little loss of generality since γ is a nuisance parameter. Then, the no sorting condition can be represented as

$$\mathbb{E}[\omega_i|W_i, J_i] = \mathbb{E}[\omega_i|W_i],$$

where for Graham W_i is the indicator of whether the person is minority. This equation imposes no systematic difference in the average of unobserved individual heterogeneity across neighborhoods. In my model, this condition corresponds to $\mathbb{E}[\omega_i|\Theta_i, W_i] = \mathbb{E}[\omega_i|W_i]$ (i.e., no selection bias condition in the above). To see this point, note $\mathbb{E}[\omega_i|W_i, J_i] = \int \mathbb{E}[\omega_i|W_i, J_i, \mathbf{A}, \Theta] f_{\Theta\mathbf{A}|W,J} = \int \mathbb{E}[\omega_i|W_i, \Theta] f_{\Theta\mathbf{A}|W,J}$, where I use (5) and (6). If ω_i and Θ_i are conditionally mean independent given W_i, the integral integrates to one, which implies $\mathbb{E}[\omega_i|W_i, J_i]$ is independent of J_i. Therefore, no selection bias in my model translates to the no sorting condition in Graham's paper. Whereas Graham explores scenarios under which this no sorting condition is plausible, I allow for failure of this condition and instead use observable shifter variables to construct control functions. □

2.2 Heuristic Discussion of Identification Result

The previous subsection indicated the source of selection bias. Here, I discuss this paper's approach for identification emphasizing main ideas rather than technical details. For simplicity, assume finite support of Θ_i, i.e., $\text{supp}(\Theta_i) = \{\theta_t : 1 \leq t \leq T\}$. We can view this discrete-valued heterogeneity Θ_i as agents' type and these types differ in preferences for group attributes A_s.

I can rewrite the outcome equation by decomposing ω_i into its mean within group J_i and the deviation.

$$Y_i = \beta' X_i + \gamma' W_i + \mathbb{E}[\omega_i|J_i, \mathbf{A}] + \varepsilon_i, \qquad \varepsilon_i = \varepsilon_{iJ_i}, \qquad \varepsilon_{is} = \omega_i - \mathbb{E}[\omega_i|J_i = s, \mathbf{A}] + \chi_s + \epsilon_{is} \quad (8)$$

where I write $X_i \equiv X_{J_i}$, $\varepsilon_i \equiv \varepsilon_{iJ_i}$ to avoid double subscripts, and $\mathbb{E}[\omega_i|J_i = s, \mathbf{A}]$ represents the

expectation of ω_i given that the individual selected group s and the underlying group features are $\mathbf{A}$. Conditioning on $\mathbf{A}$ is necessary because an agent observes and takes as given the group features $\mathbf{A}$ when making group choice.[3] In the above decomposition, X_i is uncorrelated with ε_i, and therefore, endogeneity is present if and only if X_s has non-zero correlation with the within-group mean $\mathbb{E}[\omega_i|J_i = s, \mathbf{A}]$. For this conditional expectation, we have

$$\begin{aligned}\mathbb{E}[\omega_i|J_i = s, \mathbf{A}] &= \sum_{t=1}^{T} \mathbb{E}[\omega_i|J_i = s, \mathbf{A}, \Theta_i = \theta_t] \Pr(\Theta_i = \theta_t|J_i = s, \mathbf{A}) \\ &= \sum_{t=1}^{T} \mathbb{E}[\omega_i|\mathbf{A}, \Theta_i = \theta_t] \Pr(\Theta_i = \theta_t|J_i = s, \mathbf{A}) \\ &= \sum_{t=1}^{T} \mathbb{E}[\omega_i|\Theta_i = \theta_t] \Pr(\Theta_i = \theta_t|J_i = s, \mathbf{A}). \end{aligned} \tag{9}$$

where the second equality follows from $J_i = J(\Theta_i, \mathbf{A}, \boldsymbol{\eta}_i)$ and independence $\omega_i \perp\!\!\!\perp \boldsymbol{\eta}_i|\Theta_i$, and the third equality uses $\mathbf{A} \perp\!\!\!\perp (\Theta_i, \omega_i)$.

The last equation indicates that across-group variation in $\mathbb{E}[\omega_i|J_i = s, \mathbf{A}]$ comes from variation in $\Pr(\Theta_i = \theta|J_i = s, \mathbf{A})$. The key insight of this paper is that $\Pr(\Theta_i = \theta|J_i = s, \mathbf{A})$ can be expressed as a linear function of the conditional distribution of the observable shifter W_i given $(J_i, \mathbf{A})$, and therefore, accounting for across-group variation in observable shifter distributions solves the endogeneity problem of X_s.

To sketch the identification argument, I assume finite support of W_i, i.e., $\text{supp}(W_i) = \{w_l : 1 \leq l \leq L\}$. This is without loss of generality since I can always create a discrete random variable from continuous one by binning. In the case of discretely distributed Θ_i, W_i is a shifter for Θ_i if the matrix

$$\mathbf{\Pi} = \begin{bmatrix} \Pr(\Theta_i = \theta_1|W_i = w_1) & \dots & \Pr(\Theta_i = \theta_T|W_i = w_1) \\ & \ddots & \\ \Pr(\Theta_i = \theta_1|W_i = w_L) & \dots & \Pr(\Theta_i = \theta_T|W_i = w_L) \end{bmatrix}$$

is of full column rank. This condition formalizes the idea that the conditional distribution of Θ_i given $W_i = w$ exhibits enough variation in w. Also it requires that the cardinality of $\text{supp}(W_i)$

[3] Another way to see this point is that $\mathbf{A}$ is invariant across individuals given $J_i = s$. The sample group mean is average over i, i.e., $\sum_{i=1}^{N} \mathbb{1}\{J(\Theta_i, \mathbf{A}, \boldsymbol{\eta}_i) = s\}\omega_i / \sum_{i=1}^{N} \mathbb{1}\{J(\Theta_i, \mathbf{A}, \boldsymbol{\eta}_i) = s\}$ and $\mathbf{A}$ is independent of i. Then, the sample group means converges in probability to the expecatation conditional on $\mathbf{A}$. This calculation is akin to computing probability limits of sample means over long time horizon in panel data models. Given a variable X_{it} and fixed effects α_i, $T^{-1} \sum_{t=1}^{T} X_{it} \to_{\mathbb{P}} \mathbb{E}[X_{it}|\alpha_i]$ under suitable conditions.

is larger than that of $\text{supp}(\Theta_i)$ since full column rank fails if $L < T$. This full rank condition may be reasonable when a researcher observes a rich array of individual-level variables that are related to individual taste for group-level attributes. In the neighborhood example, variables such as education level, health status, type of occupation, and ethnicity are likely to have non-trivial relationships with individual taste for various neighborhood characteristics.

Doing similar calculations to those for (9) and by Bayes' rule,

$$\begin{aligned}\Pr(W_i = w|J_i, \mathbf{A}) &= \sum_{t=1}^{T} \Pr(W_i = w|\Theta_i = \theta_t) \Pr(\Theta_i = \theta_t|J_i, \mathbf{A}) \\ &= \sum_{t=1}^{T} \Pr(\Theta_i = \theta_t|W_i = w) \Pr(W_i = w) \frac{\Pr(\Theta_i = \theta_t|J_i, \mathbf{A})}{\Pr(\Theta_i = \theta_t)}\end{aligned}$$

and dividing the both sides by $\Pr(W_i = w)$,

$$\frac{\Pr(W_i = w|J_i, \mathbf{A})}{\Pr(W_i = w)} = \sum_{t=1}^{T} \Pr(\Theta_i = \theta_t|W_i = w) \frac{\Pr(\Theta_i = \theta_t|J_i, \mathbf{A})}{\Pr(\Theta_i = \theta_t)}$$

We can write this equation in matrix form

$$\boldsymbol{\pi}_w = \boldsymbol{\Pi}\boldsymbol{\pi}_\theta \tag{10}$$

where $\boldsymbol{\pi}_w = [\Pr(W_i = w_l|J_i, \mathbf{A})/\Pr(W_i = w_l)]_{l=1}^{L}$ and $\boldsymbol{\pi}_\theta = [\Pr(\Theta_i = \theta_t|J_i, \mathbf{A})/\Pr(\Theta_i = \theta_t)]_{t=1}^{T}$. Since $\boldsymbol{\Pi}$ is of full column rank by the observable shifter assumption, we have $\boldsymbol{\pi}_\theta = \mathbf{G}\boldsymbol{\pi}_w$ where $\mathbf{G} = (\boldsymbol{\Pi}'\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}'$, and the t-th element of this linear equation looks like

$$\Pr(\Theta_i = \theta_t|J_i, \mathbf{A}) = \Pr(\Theta_i = \theta_t) \sum_{l=1}^{L} G_{tl} \frac{\Pr(W_i = w_l|J_i, \mathbf{A})}{\Pr(W_i = w_l)} \equiv \sum_{l=1}^{L} d_{tl} \Pr(W_i = w_l|J_i, \mathbf{A})$$

where G_{tl} is the (t, l)th element of $\mathbf{G}$ and d_{tl} is a constant independent of $(J_i, \mathbf{A})$. Substituting this last result into (9), we obtain

$$\begin{aligned}\mathbb{E}[\omega_i|J_i, \mathbf{A}] &= \sum_{t=1}^{T} \mathbb{E}[\omega_i|\Theta_i = \theta_t] \sum_{l=1}^{L} d_{tl} \Pr(W_i = w_l|J_i, \mathbf{A}) \\ &= \sum_{l=1}^{L} \delta_l \Pr(W_i = w_l|J_i, \mathbf{A}), \qquad \delta_l = \sum_{t=1}^{T} d_{tl}\mathbb{E}[\omega_i|\Theta_i = \theta_t]\end{aligned}$$

and finally the outcome equation becomes

$$Y_i = \beta' X_i + \gamma' W_i + \sum_{l=1}^{L} \delta_l \Pr(W_i = w_l | J_i, \mathbf{A}) + \varepsilon_i$$

where $\Pr(W_i = w_l | J_i, \mathbf{A})$ is identifiable from the data and $\text{Cov}(X_i, \varepsilon_i) = 0$. Therefore, the vector $[\Pr(W_i = w_l | J_i, \mathbf{A})]_{l=1}^{L}$ plays the role of a control function to account for the endogeneity of X_s with respect to unobserved individual heterogeneity ω_i.

3 Identification Result

In this section, I formalize the heuristics described in the preceding section. For that purpose, I impose the following conditions on the model. Below, let $L^p(\mu)$ be the class of functions whose pth power is integrable with respect to μ. For a random object V, I write F_V for its distribution function.

Assumption 1. *Let* $\mathbf{B}_g \equiv (\mathbf{X}_g, \boldsymbol{\chi}_g, \mathbf{A}_g)$.

(i) $\mathbf{B}_g$ *is a random draw across* g. *Conditional on* $\mathbf{B}_g$, $(W_{ig}, \Theta_{ig}, \omega_{ig}, \boldsymbol{\epsilon}_{ig}, \boldsymbol{\eta}_{ig})$ *is i.i.d. across* i, *and* $(W_{ig}, \Theta_{ig}, \omega_{ig}, \boldsymbol{\epsilon}_{ig}, \boldsymbol{\eta}_{ig})$ *is independent across* g.

(ii) $(W_{ig}, \omega_{ig}, \Theta_{ig}) \perp\!\!\!\perp \mathbf{B}_g$ *and* $\boldsymbol{\eta}_{ig} \perp\!\!\!\perp (W_{ig}, \omega_{ig}) | \Theta_{ig}, \mathbf{B}_g$.

Assumption 2. *The distribution of* $\mathbf{B}$ *is dominated by some measure. The random vectors* (W_i, Θ_i) *have a joint density with respect to the product measure of some* σ*-finite measures* μ *and* λ. *The densities* $f_{\Theta|W}, f_\Theta, f_W$ *are bounded.*

Assumption 3. *The following mapping* $\boldsymbol{\Psi}$ *defined on* $L^2(F_\Theta)$ *is injective;*

$$(\boldsymbol{\Psi} h)(w) = \int h(\theta) f_{\Theta|W}(\theta|w) d\lambda(\theta).$$

Also, $\{f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b})/f_\Theta(\theta) : s \in \mathcal{S}, \mathbf{b} \in \text{supp}(\mathbf{B})\} \subset L^2(F_\Theta)$.

From this assumption, Theorem 15.16 in Kress (2014) implies that there exist $\{\tau_j \geq 0 : j \in \mathbb{N}\}$ and an orthonormal basis $\{\phi_j : j \in \mathbb{N}\}$ on $L^2(F_\Theta)$ such that $\tau_j^2 \phi_j = \boldsymbol{\Psi}^* \boldsymbol{\Psi} \phi_j$ where $\boldsymbol{\Psi}^*$ is the adjoint operator of $\boldsymbol{\Psi}$. Assume τ_j is ordered such that $\tau_j \geq \tau_{j+1}$ for all $j \geq 1$.

Assumption 4. *With* $\{(\tau_j, \phi_j) : j \in \mathbb{N}\}$ *defined above,*

$$f_{\Theta|J,\mathbf{B}}(\cdot|s,\mathbf{b})/f_\Theta(\cdot) \in \mathcal{F}_1 = \left\{ f \in L^2(F_\Theta) : \sum_{j=1}^{M} \left|\langle f, \phi_j \rangle_\Theta\right| < \infty \right\} \text{ for all } s \in \mathcal{S}, \mathbf{b} \in \text{supp}(\mathbf{B})$$

$$\mathbb{E}[\omega_i|\Theta_i = \cdot] \in \mathcal{F}_2 = \left\{ f \in L^2(F_\Theta) : \sum_{j=1}^{M} \tau_j^{-1} \left|\langle f, \phi_j \rangle_\Theta\right| < \infty \right\},$$

where $M = \sup\{j : \tau_j > 0\}$ *and* $\langle f, g \rangle_\Theta = \int f g dF_\Theta$. M *can be positive infinity.*

Assumption 1 describes the sampling. Part (i) formalizes the idea that the distribution of group-level variables and the within-group distribution of individual observables are identifiable from the data. Part (ii) restates conditions (5) and (6) in a weaker form. Assumption 2 imposes mild restrictions on the distribution of $(\Theta_i, W_i, \mathbf{B})$.

Assumption 3 is a generalization of the full column rank condition used in the previous section. It ensures that the integral equation, which is the infinite-dimensional analogue of (10), is "invertible" in a suitable sense. This injectivity condition, usually referred to as L^2-completeness, is a high-level assumption but has been widely used in the recent econometric literature on nonparametric identification (see Section 1.1). Intuitively, injectivity requires the density of $f_{\Theta|W}(\cdot|w)$ to sufficiently vary in the conditioning value w. One example of (Θ, W) satisfying completeness is $\Theta = \Gamma W + \nu$ where ν given W is distributed as Normal$(0, \Sigma)$, Σ is invertible, and Γ is of full column rank. More sufficient conditions for different types of completeness can be found in the literature (Andrews, 2017; D'Haultfoeuille, 2011; Hu, Schennach, and Shiu, 2017; Hu and Shiu, 2018; Mattner, 1993). One way to justify this technical condition is to assume that in the population, there are a finite number of agent types which differ in preferences. Under this assumption, injectivity reduces to the full column rank of the conditional probability matrix as analyzed in the previous section. The strategy of modeling unobserved heterogeneity as finite number of types has been employed in empirical studies. For instance, recent papers of Abowd, McKinney, and Schmutte (2018) and Bonhomme, Lamadon, and Manresa (2018) study the consequence of worker-firm matching on labor market earnings and model worker and/or firm heterogeneity as discrete types. Viewing unobserved heterogeneity as finite types provides one way to rationalize Assumption 3. Nonetheless, completeness is applicable more generally and the condition can hold with continuous heterogeneity variable as well.

Also, Assumption 3 imposes restrictions on the relative tail of $f_{\Theta|J,\mathbf{B}}$ and f_Θ. In this model, selection into groups allows the within-group distribution of Θ to differ from the original distribution of Θ to the extent that this tail condition is satisfied. For instance, the model excludes the case where Θ has a Gaussian tail and the within-group distribution has a polynomial tail. This is a high-level condition since it is not straightforward to characterize the within-group distribution from the model primitives. Also, note the conditioning on $\mathbf{B}$. In Section 2, I impose stronger independence assumptions and only condition on $\mathbf{A}$, but here I slightly weaken the independence assumptions and thus conditioning on $\mathbf{B}$ is necessary.

Assumption 4 restricts the permissible classes for $f_{W|J,\mathbf{B}}$ and $\mathbb{E}[\omega_i|\Theta_i]$. It requires that

$$\begin{aligned} f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b})/f_\Theta(\theta) &= \sum_{j=1}^{\infty} c_j(s,\mathbf{b})\phi_j(\theta), \qquad & \sum_{j=1}^{\infty} |c_j(s,\mathbf{b})| < \infty, \\ \mathbb{E}[\omega_i|\Theta_i = \theta] &= \sum_{j=1}^{\infty} d_j\phi_j(\theta), & \sum_{j=1}^{\infty} |d_j/\tau_j| < \infty, \end{aligned}$$

for some $c_j(s,\mathbf{b}) \in \mathbb{R}$, $d_j \in \mathbb{R}$, $j \in \mathbb{N}$. A restrictive, yet easily interpretable sufficient condition is that there exist integers M and L, possibly dependent on $(s,\mathbf{b})$, such that $c_j(s,\mathbf{b}) = 0$ for all $j > L$ and $d_j = 0$ for all $j > M$. This condition implies that $f_{\Theta|J,\mathbf{B}}/f_\Theta$ and $\mathbb{E}[\omega_i|\Theta_i = \cdot]$ can be represented as finite linear combinations of the L^2-basis $\{\phi_j : j \in \mathbb{N}\}$.

Now, I state a key lemma for the identification of group-level partial effects.

Lemma 1. *Suppose Assumptions 1-4 hold. Then, there exists some function ψ such that*

$$\mathbb{E}[\omega_i|J_i = s, \mathbf{B}] = \int \psi(w) dF_{W|J,\mathbf{B}}(w|s,\mathbf{B})$$

where $F_{W|J,\mathbf{B}}$ is the conditional distribution function of W_i given $J_i, \mathbf{B}$ and $\int \psi^2 dF_W < \infty$.

This lemma suggests a way to control for the selection bias. To see the implication, from (8),

$$\begin{aligned} Y_i &= \beta' X_i + \gamma' W_i + \int \psi(w) dF_{W|J,\mathbf{B}}(w|J_i,\mathbf{B}) + \{\omega_i - \mathbb{E}[\omega_i|J_i,\mathbf{B}]\} + \chi_i + \epsilon_i \\ &= \beta' X_i + \gamma' W_i + \sum_{k=1}^{\infty} \delta_k \mathbb{E}[p_k(W_i)|J_i,\mathbf{B}] + \varepsilon_i \end{aligned} \tag{11}$$

where $\chi_i = \chi_{J_i}$, $\epsilon_i = \epsilon_{iJ_i}$, $\varepsilon_i = \varepsilon_{iJ_i}$, $\varepsilon_{is} = \omega_i - \mathbb{E}[\omega_i|J_i,\mathbf{B}] + \chi_s + \epsilon_{is}$, $\{p_k : k \in \mathbb{N}\}$ is a basis

for $L^2(F_W)$, and $\{\delta_k : k \in \mathbb{N}\}$ satisfies $\int(\psi - \sum_{k=1}^{K} \delta_k p_k)^2 dF_W \to 0$ as $K \to \infty$. The class of approximating functions is chosen by the researcher and the conditional distribution of W_i given J_i and $\mathbf{B}$ is just a within group distribution of W_i, which is identifiable from the data. Since ε_i is uncorrelated from X_i, inclusion of $\{\mathbb{E}[p_k(W_i)|J_i, \mathbf{B}] : k \in \mathbb{N}\}$ controls for endogeneity arising from selection.

Example 2 (Continued)**.** Altonji and Mansfield (2018) also use a control function approach to address selection bias under a very similar econometric model. Despite some differences in imposed conditions,[4] Lemma 1 generalizes their result. In particular, Proposition 1 in Altonji and Mansfield states

$$\mathbb{E}[\omega_i|J_i, \mathbf{B}] = \pi' \mathbb{E}[W_i|J_i, \mathbf{B}]$$

for some π. By taking $\psi(w) = \pi' w$, Lemma 1 in this paper and their Proposition 1 coincide.

Lemma 1 encompassing their Proposition 1 has an important implication for estimation. The estimation method considered in this paper approximates ψ using a series basis expansion. Suppose a researcher chooses polynomial or spline basis functions. Then, the estimation equation based on Lemma 1 specializes to the version used by Altonji and Mansfield if series expansion terminates after the constant and linear terms. Therefore, we can view the series-based estimation proposed in this paper as a robustified version of Altonji and Mansfield's control function method. □

3.1 Main Result

Now I formally state the identification of group-level partial effects. The object $F_{W|J,\mathbf{B}}(\cdot|J_i, \mathbf{B})$ is a random distribution function, i.e., a function-valued random element. Thus, there exists a σ-field generated by $F_{W|J,\mathbf{B}}(\cdot|J_i, \mathbf{B})$ and the conditional expectation given $F_{W|J,\mathbf{B}}(\cdot|J_i, \mathbf{B})$ is well defined.

Assumption 5. *(i) For* $s \in \mathcal{S}$, $\mathbb{E}[\varepsilon_{ig}|J_{ig} = s, F_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)] = 0$ *and* $\mathbb{E}[V_{ig}\varepsilon_{ig}] = 0$ *where* $\varepsilon_{ig} = \varepsilon_{iJ_{ig}g}$ *and* $V_{ig} = (X_{ig}' \, W_{ig}')'$.

(ii) The matrix

$$\mathbb{E}[\{V_{ig} - \mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)]\}\{V_{ig} - \mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)]\}']$$

[4] I elaborate on different assumptions between the two papers in Section 5.1.1.

is invertible.

Assumption 5 (i) imposes exogeneity of χ_s and excludes the case of omitted variable bias. It also assumes that W_i is uncorrelated with $\{\omega_i - \mathbb{E}[\omega_i|J_i, \mathbf{B}]\}$. Since γ is not the parameter of interest, I can rewrite the regression equation by projecting $\{\omega_i - \mathbb{E}[\omega_i|J_i, \mathbf{B}]\}$ onto W_i to make them uncorrelated. Part (ii) is a version of the key identification condition for partially linear models. In the regression equation (11), the interest lies in β and we want to "partial out" the nuisance parameter $\int \psi dF_{W|J,\mathbf{B}}$. Here the within-group distribution of W_i, denoted by $F_{W|J,\mathbf{B}}$, is estimable from the data and conditional on the value of $F_{W|J,\mathbf{B}}$, $\int \psi dF_{W|J,\mathbf{B}}$ is a constant and thus differencing eliminates this nuisance parameter.

Part (ii) requires that X_s does not include functions of the within-group distribution of W_i, e.g., the within-group mean of W_i. This follows from $\mathbb{E}[h(W_i)|F_{W|J,\mathbf{B}}(\cdot|J_i, \mathbf{B})] = \mathbb{E}[h(W_i)|J_i, \mathbf{B}]$ for every measurable function h.[5] Sometimes it is desirable to include within-group distributional features in X_s since they represent "peer effects" as in Example 1. To accommodate such situations, it suffices to have a subvector $W_i^{\mathtt{sub}}$ of W_i that satisfies Assumption 3, i.e., $(\Theta_i, W_i^{\mathtt{sub}})$ is L^2-complete. Then, a researcher can include distributional features of W_i in X_s provided that the included elements are not part of $W_i^{\mathtt{sub}}$.

Building on Lemma 1, the following theorem formalizes the identification of coefficients on group-level variables.

Theorem 1. *Suppose that Assumptions 1-5 hold. Then, β in the equation (2) is identified.*

4 Estimation and Inference

Lemma 1 suggests an estimation method by series approximation. If a researcher observed within-group distributions of W_{ig}, then the estimator based on least squares is

$$\hat{\beta}_{\text{oracle}} = \mathbf{S}(\mathbf{P}_K'\mathbf{P}_K)^{-1}\mathbf{P}_K\mathbf{Y}$$

[5]This claim follows if $\mathbb{E}[h(W_i)|J_i, \mathbf{B}]$ is measurable with respect to $\sigma(F_{W|J,\mathbf{B}})$, the σ-field generated by $F_{W|J,\mathbf{B}}(\cdot|J_i, \mathbf{B})$. A σ-field on the space of distribution functions can be defined by the σ-field generated by maps $h \to \int h dF$ where $h \geq 0$ is a measurable function from $\mathbb{R}^k \to \mathbb{R}$ and F is a random distribution function (see e.g., Kallenberg, 2017). The measurability of $\mathbb{E}[h(W_i)|J_i, \mathbf{B}]$ follows from the representation $\mathbb{E}[h(W_i)|J_i, \mathbf{B}] = \int h(w) dF_{W|J,\mathbf{B}}(w|J_i, \mathbf{B})$.

where $\mathbf{P}_K = [P_{K11}P_{K21}\dots P_{KNG}]'$, $P_{Kig} = (X_{ig}', W_{ig}', \mathbb{E}[p_1(W_{ig})|J_{ig}, \mathbf{B}_g], \dots, \mathbb{E}[p_K(W_{ig})|J_{ig}, \mathbf{B}_g])'$, K represents the number of series terms, $\mathbf{Y} = (Y_{11}, \dots, Y_{NG})'$, and $\mathbf{S} = [\mathrm{I}_{d_x} \mathbf{O}_{d_x \times (d_x + d_w + K)}]$ where d_x, d_w denote the dimensions of X, W, respectively, I_d is the $d \times d$ identity matrix, and $\mathbf{O}_{d_1 \times d_2}$ is the $d_1 \times d_2$ matrix with all elements equal to zero.

In practice, a researcher needs to estimate within-group distributions. Define the following object

$$\hat{\mathbb{E}}[p_k(W_{ig})|J_{ig}, \mathbf{B}_g] = \sum_{j=1}^{N} p_k(W_{jg}) \mathbb{1}\{J_{jg} = J_{ig}\} \Big/ \sum_{j=1}^{N} \mathbb{1}\{J_{jg} = J_{ig}\}.$$

Then, the feasible version of the above estimator is

$$\hat{\beta} = \mathbf{S}(\hat{\mathbf{P}}_K' \hat{\mathbf{P}}_K)^{-1} \hat{\mathbf{P}}_K \mathbf{Y}$$

where $\hat{\mathbf{P}}_K = [\hat{P}_{K11} \ \dots \ \hat{P}_{KNG}]'$ and $\hat{P}_{Kig} = (X_{ig}', W_{ig}', \hat{\mathbb{E}}[p_1(W_{ig})|J_{ig}, \mathbf{B}_g], \dots, \hat{\mathbb{E}}[p_K(W_{ig})|J_{ig}, \mathbf{B}_g])'$.

To analyze asymptotic properties of $\hat{\beta}$, I impose additional assumptions. To state the conditions, write $\lambda_{\max}(\cdot), \lambda_{\min}(\cdot)$ for the maximum and minimum of eigenvalues of the argument. Let $\|\cdot\|$ be the Euclidean norm for vectors and the induced norm for matrices. Also, write $\varepsilon_{ig} = \varepsilon_{iJ_{ig}g}$ to avoid double subscripts.

Assumption 6. *The basis functions $\{p_k : k \in \mathbb{N}\}$ are uniformly bounded and the eigenvalues of the matrix $\mathbf{Q}_K = \mathbb{E}[P_{Kig}P_{Kig}']$ are bounded and bounded away from zero uniformly in K. In addition, the class $\{p_k : k \in \mathbb{N}\}$ is a subset of some VC-class of functions as defined in van der Vaart and Wellner (1996). In addition, for some $m_x, m_w > 4$, $\mathbb{E}[\|W_{ig}\|^{m_w}] + \mathbb{E}[\|X_{sg}\|^{m_x}] < \infty$ for all $s \in \mathcal{S}$.*

Assumption 7. *Write $\mathcal{A}_{ig}$ for the σ-field generated by $\{W_{ig}, X_{ig}, J_{ig}, \mathbb{E}[p_k(W_{ig})|J_{ig}, \mathbf{B}_g]\ k \in \mathbb{N}\}$. Then, $\mathbb{E}[\varepsilon_{ig}|\mathcal{A}_{ig}] = 0$, and there exist fixed constants c_1, c_2, C_1, C_2 such that, with probability one, $0 < c_1 \leq \mathbb{E}[\varepsilon_{ig}\varepsilon_{jg}|\mathcal{A}_{ig}, \mathcal{A}_{jg}] \leq C_1$, $\mathbb{E}[\varepsilon_{ig}^4|\mathcal{A}_{ig}, J_{ig}] \leq C_2$, and $\lambda_{\min}(\mathbb{E}[P_{Kig}P_{Kjg}'\varepsilon_{ig}\varepsilon_{jg}]) \geq c_2 > 0$ for all K.*

Assumption 8. *Let $r_{Kig} = \sum_{k=K+1}^{\infty} \delta_k \mathbb{E}[p_k(W_{ig})|J_{ig}, \mathbf{B}_g]$. There exists some $b > 0$ such that $\mathbb{E}[r_{Kig}^2] = O(K^{-2b})$.*

The above assumptions are relatively standard in the series estimation literature. However, the "basis functions" used in this procedure are conditional expectations of basis functions. Therefore,

there are no general sufficient conditions for $\lambda_{\max}(\mathbf{Q}_K)$ to be bounded above and for $\lambda_{\min}(\mathbf{Q}_K)$ and $\lambda_{\min}(\mathbb{E}[P_{Kig}P'_{Kjg}\varepsilon_{ig}\varepsilon_{jg}])$ to be bonded away from zero. Since these high-level conditions may not be easily verifiable, one approach is to take a "flexible parametric" view, which is to assume that the function ψ in Lemma 1 can be approximated by finitely many but unknown number of basis functions. In that case, the model is essentially parametric and verifications of the conditions become straightforward.

The boundedness of basis functions in Assumption 6 holds for most of basis functions used in practice if the support of W_{ig} is compact. The unbounded support case would require use of different norms. For its second part, I assume that the second moment matrix of P_{Kig} is positive definite for all K and, in particular, that the eigenvalues are bounded above and away from zero. However, as mentioned above, there are no general primitive conditions for bounded eigenvalues as the series terms are the conditional expectations of original basis functions. In the proof, I make it explicit how the estimator depends on the minimum/maximum eigenvalues to understand how eigenvalues tending to zero/infinity may affect the asymptotic distributional properties of the estimator.

Assumption 7 imposes boundedness of the conditional variance of ε_{isg} from above and below and boundedness of the conditional fourth moment. The restriction on the conditional variance seems to be standard in the literature and I use the finite fourth moment to verify Lindberg condition for central limit theorem. Assumption 8 controls the bias term in series approximation. This condition can be verified for specific $\{p_k : k \in \mathbb{N}\}$ if the function ψ has enough smoothness (see e.g., Belloni, Chernozhukov, Chetverikov, and Kato, 2015, and references therein).

Now, I formally state the asymptotic properties of $\hat{\beta}$. Define $\xi_K = \sup_w \|(p_1(w), \ldots, p_K(w))'\|$, and $\zeta_K = \xi_K + (NG)^{1/m_w} + G^{1/m_x}$ where m_w, m_x are defined in Assumption 6. In practice, this quantity ζ_K is at least as large as $\sqrt{K}$, and below I take $\zeta_K \geq \sqrt{K}$.

Theorem 2. *Suppose β is identified in (11) and for some $c > 0$, $\Pr(N_{sg}/N_g \geq c$ for all $s \in \mathcal{S}$ and $g) \to 1$ where $N_{sg} = \sum_{i=1}^N \mathbb{1}\{J_{ig} = s\}$ and $N_g = \sum_{s=1}^S N_{sg}$. Under Assumptions 1, 6-8, if $\zeta_K^2 \log K/G \to 0$ and $G/K^{2b} = o(1)$, then*

$$\sqrt{G}(\hat{\beta}_{oracle} - \beta) = \mathbf{S}\mathbf{Q}_K^{-1} \frac{1}{\sqrt{G}} \sum_{g=1}^G \psi_{Ng} + o_{\mathbb{P}}(1)$$

where

$$\psi_{Ng} = \frac{1}{N}\sum_{i=1}^{N} P_{Kig}\varepsilon_{ig}.$$

Furthermore, if $\mathrm{Var}(\chi_{sg} - \mathbb{E}[\omega_{ig}|J_{ig} = s, \mathbf{B}_g]) > 0$ *for each* s *and* $\zeta_K^4/G \to 0$*, then*

$$\mathbf{\Omega}_K^{-1/2}\mathbf{S}\mathbf{Q}_K^{-1}\frac{1}{\sqrt{G}}\sum_{g=1}^{G}\psi_{Ng} \rightsquigarrow \mathrm{Normal}(\mathbf{0}, \mathrm{I}_{d_x})$$

where

$$\mathbf{\Omega}_K = \mathbf{S}\mathbf{Q}_K^{-1}\mathbb{E}[\psi_{Ng}\psi_{Ng}']\mathbf{Q}_K^{-1}\mathbf{S}'$$

and $\|\mathbf{\Omega}_K^{-1}\|$ *is uniformly bounded. Finally, if* $\zeta_K\sqrt{K\log(NG)/N} \to 0$*, then*

$$\sqrt{G}(\hat{\beta} - \hat{\beta}_{oracle}) = o_{\mathbb{P}}(1).$$

Theorem 2 characterizes asymptotic distribution of $\hat{\beta}$ and provides a set of sufficient conditions under which $\hat{\beta}$ is asymptotically equivalent to the oracle estimator $\hat{\beta}_{\text{oracle}}$. It also states that the convergence rate of $\hat{\beta}$ is $\sqrt{G}$ rather than $\sqrt{GN}$. That N does not show up in the convergence rate comes from within-group correlations of residuals. The residual ε_{isg} contains group-level unobservable χ_{sg}, which causes within-group dependence. The within-group correlation is formalized by the requirement $\mathrm{Var}(\chi_{sg} - \mathbb{E}[\omega_{ig}|J_{ig} = s, \mathbf{B}_g]) > 0$. As seen below, this within-group correlation requires the standard error estimator clustered at group levels.

In the first part of the theorem, I assume that the number of individuals in each group grows at a proportional rate with the city-wide number of individuals for every city. This requires that the probability of selecting into a group is uniformly bounded away from zero for all groups. This assumption guarantees that I can estimate the within-group means of $p_k(W_{ig})$ uniformly well across groups and cities.

In the theorem, I also impose restrictions on (relative) growth rates of K, N, and G. The first two are $\zeta_K^2 \log K/N = o(1)$ and $G/K^{2b} \to 0$. These conditions are standard in the literature except that I have ζ_K^2, which includes terms related to the number of finite moments of W_{ig} and X_{sg}. The non-standard part of the requirement is $\zeta_K^4/G \to 0$. This condition requires at least fourth moments of W_{ig} and X_{sg} to be finite. Particularly, it implies that $\mathbb{E}[\|X_{sg}\|^{m_x}] < \infty$ with $m_x > 4$ and

$\mathbb{E}[\|W_{ig}\|^{m_w}] < \infty$ with $m_w > 4$ satisfying $N^4/G^{(m_w-4)} \to 0$. The finite moment condition allows control of the rate at which the maximum of $\|W_{ig}\|$ and $\|X_{ig}\|$ over (i,g) grows. In addition, this rate condition requires that $\xi_K^4/G \to 0$, which is stronger than what is used in the literature (e.g., Cattaneo, Farrell, and Feng, 2018). Since this paper's emphasis is on the constructive identification result, I maintain this assumption and plan to improve on this aspect of the theoretical result in the future research.

For inference, we need a consistent estimator of the variance. A natural estimator of $\mathbf{\Omega}_K$ is

$$\hat{\mathbf{\Omega}}_K = \mathbf{S}\tilde{\mathbf{Q}}_K^{-1}\hat{\mathbf{\Sigma}}_K\tilde{\mathbf{Q}}_K^{-1}\mathbf{S}'$$

where $\tilde{\mathbf{Q}}_K = \hat{\mathbf{P}}_K'\hat{\mathbf{P}}_K/GN$, $\hat{\mathbf{\Sigma}}_K = \frac{1}{G}\sum_{g=1}^{G}\hat{\psi}_{Ng}\hat{\psi}_{Ng}'$, $\hat{\psi}_{Ng} = \frac{1}{N}\sum_{i=1}^{N}\hat{P}_{Kig}(Y_{ig} - \hat{P}_{Kig}'\hat{\theta})$, and $\hat{\theta} = (\hat{\mathbf{P}}_K'\hat{\mathbf{P}}_K)^{-1}(\hat{\mathbf{P}}_K'\mathbf{Y})$. The following theorem formalizes that this variance estimator is consistent.

Theorem 3. *In addition to the hypothesis of Theorem 2, if* $\zeta_K^3\sqrt{K/G} + \zeta_K^3 K^{-b} \to 0$ *and for some* $m_\varepsilon > 2$, $\mathbb{E}[|\varepsilon_{isg}|^{m_\varepsilon}] < \infty$ *and* $\zeta_K\sqrt{(GN)^{1/m_\varepsilon}\log K/G} \to 0$ *hold, then*

$$\|\hat{\mathbf{\Omega}}_K^{-1/2} - \mathbf{\Omega}_K^{-1/2}\| = o_{\mathbb{P}}(1).$$

5 Numerical Results

5.1 Empirical Application

I employ the results of this paper to study the neighborhood/school-district effects on student outcomes. Particularly, I use the National Longitudinal Study of 1972 (NLS72), which was one of the datasets analyzed by Altonji and Mansfield (2018).[6] They consider an econometric model very similar to (2)-(3). However, the selection equation in this paper is more general and I impose a set of different conditions to achieve identification. Their Proposition 1 implies that the outcome equation can be written as

$$Y_{is} = X_s'\beta + W_i'\gamma + \pi'\mathbb{E}[W_i|J_i,\mathbf{B}] + \varepsilon_{is}$$

[6] They analyze three other datasets in their paper, all of which are restricted-use. The dataset based on NLS72 is publicly available.

and the term $\pi'\mathbb{E}[W_i|J_i, \mathbf{B}]$ plays the role of a control function to eliminate selection bias. Lemma 1 in this paper suggests that in general the control function may not be linear in within-group means of observable shifter W_i. Since least squares estimation based on series expansions covers the general nonlinear case and nests the linearity case, the estimation method proposed in this paper can test whether the linear specification is a reasonable approximation of the data generating process.

An empirically relevant concern is that even after including control functions, there remains "omitted variable bias" because X_s and χ_s are potentially correlated and a researcher does not observe χ_s. That is, the least squares estimate of β converges in probability to $\tilde{\beta} = \beta + \mathbb{E}[X_s X_s']^{-1}\mathbb{E}[X_s \chi_s]$, which includes the coefficient of linear projection of χ_s on X_s. Altonji and Mansfield address this concern by developing lower bounds for some measures of group-level effects on outcomes. They look at the impact of shifting from the 10th to the 90th quantile of the school/neighborhood characteristics, $\mathrm{Q}_{90}(X_s'\beta + \chi_s) - \mathrm{Q}_{10}(X_s'\beta + \chi_s)$. If $X_s'\beta + \chi_s$ is normally distributed, this difference in quantile can be expressed as $2 * 1.28 * \sqrt{\mathrm{Var}(X_s'\beta + \chi_s)}$, and Altonji and Mansfield provide conditions under which $\mathrm{Var}(X_s'\tilde{\beta}) \leq \mathrm{Var}(X_s'\beta + \chi_s)$. Since $\mathrm{Var}(X_s'\tilde{\beta})$ is estimable from the data, this approach produces feasible lower bounds for the object of interest.[7]

I follow the same lower bound approach to study the school/neighborhood contribution to early adulthood wage earnings of students and years of post-secondary education. Table 1 displays estimates of measures of school/neighborhood effects on early adulthood log wage (with two sets of regressors) and years of post-secondary education. For each outcome variable, I estimated the regression model in two specifications. The first one uses group means of individual covariates, which is the specification of Altonji and Mansfield, and the second contains group means of interaction and squared terms as additional controls. In the table, "AM" denotes the first specification and "This Paper" refers to the second. The first and second rows are the impact of moving a student from a 10th percentile neighborhood to a 90th/50th percentile one i.e., $\{\Phi^{-1}(q) - \Phi^{-1}(0.1)\}\sqrt{\mathrm{Var}(X_s'\beta)}$, $q \in \{0.9, 0.5\}$ where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function.

Across outcome variables and estimands, the linear and quadratic specifications produce very similar estimates. For the log wage outcome, moving a student from a 10th percentile school to

[7] Altonji and Mansfield also look at the fraction of variance attributable to school/neighborhood quality. They employ random effects modeling to estimate the variance of the school-level unobserved term and decompose the variance to estimate the (lower bound of the) variance fraction.

a 90th percentile increases the wage by 17.58% $\approx (100 * \exp(0.162) - 100)$ with post-secondary education as a control and by 17.11% without the control. For the 10th versus the 50th percentile, the increases are 8.44% and 8.22% for the two specifications. For years of post-secondary education, the shift from a 10th to 90th/50th percentile school induces increases of 0.37 and 0.19 years, which correspond to 0.22 and 0.11 standard deviation, respectively. These estimates suggest non-trivial effects of school/neighborhood quality on how many years of education students attain after graduating from high school.

The estimates in this paper differ from those in Altonji and Mansfield to some degree. For instance, they report the point estimates of 0.121 and 0.125 for log wage increase by moving from the 10th percentile to the 90th percentile, which contrast with 0.158 and 0.162 in this paper. The discrepancy arises because Altonji and Mansfield use a random effects model to estimate the variance of group-level unobservable term whereas I use the ordinary least squares (OLS) method. When I employ the same estimation approach, the quadratic specification still produces estimates very similar to the ones based on the specification of Altonji and Mansfield (see Table 3). I choose the OLS results as my main estimates for consistency with the theoretical results proven in this paper. However, one can expect that under appropriate conditions, methods based on random effects models will be valid using the control function method.

5.1.1 Differences in Assumptions

In this subsection, I discuss the main differences between the assumptions in this paper and those of Altonji and Mansfield (2018). Their key conditions include linearity of the regression functions $\mathbb{E}[W_{ig}|\Theta_{ig}], \mathbb{E}[\omega_{ig}|\Theta_{ig}]$ and what they term "spanning assumption." To define the spanning assumption, introduce the new random vector U_{ig} and write

$$\omega_{ig} = c' U_{ig}$$
$$\Theta_{ig} = \Gamma W_{ig} + \Delta U_{ig} + \nu_{ig}$$

where $c \in \mathbb{R}^{d_u}$, $\Gamma \in \mathbb{R}^{d_\theta \times d_w}$, and $\Delta \in \mathbb{R}^{d_\theta + d_u}$ are fixed parameters, and $\nu_{ig} \in \mathbb{R}^{d_\theta}$ is a random vector uncorrelated with $(W_{ig}', U_{ig}')'$. Let Π_{UW} be the matrix of coefficients for linear projection of U_{ig} onto W_{ig}. Then, the spanning assumption of Altonji and Mansfield is that there exists a matrix R

such that

$$\Delta = (\Gamma + \Delta\Pi_{UW})R.$$

Intuitively, the spanning assumption requires that for each element of A_s, if U_i influences the taste coefficient for that element of A_s, then either one of them has to hold: W_i affects the taste coefficient directly by Γ having non-zero elements or indirectly by non-zero correlation with U_i.

In this paper, I use an injectivity condition (Assumption 3) as well as restrictions on permissible classes of densities and conditional expectations to construct a control function. Unlike Altonji and Mansfield, I do not impose linearity in conditional expectations and instead use restrictions on the distribution, which makes the spanning assumption and Assumption 3 quite different. In fact, there are examples where one of the two holds but the other fails. However, if $(W_{ig}', U_{ig}', \nu_{ig}')'$ are jointly normal, the two conditions coincide. Despite the differences, Lemma 1 nests Proposition 1 in Altonji and Mansfield and therefore, the estimator proposed in this paper is generally more robust than the one in their paper.

5.2 Monte Carlo Experiments

In this subsection, I present results of simulation studies to investigate finite-sample properties of the proposed estimator. For the Monte Carlo experiments, I first generate characteristics of individuals and groups based on distributions mimicking the empirical distributions of the NLS72 data. Then, for each individual, I compute utility functions of choosing different groups and then assign people to groups to maximize the sum of utilities under group size constraints. This design builds on Altonji and Mansfield, and the realized allocation can be viewed as an approximation to the equilibrium of a competitive market through price mediation.

I consider three data generating processes (DGPs). The first design has $\mathbb{E}[\omega_i|J_i, \mathbf{A}] = \pi'\mathbb{E}[W_i|J_i, \mathbf{A}]$ for some non-stochastic vector π, and thus the linear specification of Altonji and Mansfield is correct. For the second DGP, I take ω_i to be a linear combination of some normal random variable and binary variable $\mathbb{1}\{W_i'\tau > \nu_i\}$ where τ is a non-stochastic vector and ν_i is an independent normal random variable. This design is meant to capture a more realistic scenario in which non-linearity may be a issue. Here non-linearity arises due to discreteness of part of ω_i, though ω_i itself is continuously distributed. For the third DGP, I model $\mathbb{E}[\omega_i|J_i, \mathbf{A}] = \pi_1\mathbb{E}[W_i^2|J_i, \mathbf{A}] + \pi_2\mathbb{E}[W_i^3|J_i, \mathbf{A}]$ to

see how deviation from linearity affects the estimator. See Section B in the appendix for further details of the DGPs.

Table 2 presents the mean squared error (MSE) and the coverage probability of the 95% confidence intervals based on the proposed estimator for three different specifications: one with perfect control on selection bias (i.e., "oracle" estimator), the linear specification of Altonji and Mansfield, and the one including additional series approximation terms. For DGP 1, linear specification is correct and including additional control variables does not alter the results much. For the second DGP, the group-level expectation of ω_i is $\Phi(W_i'\tau)$ where $\Phi(\cdot)$ is the standard normal cumulative distribution function, and both specifications are mis-specified. Yet, the estimator with cubic splines controls outperforms the linear one. For the third design, I made the DGP particularly difficult for the linear specification to see how deviation from linearity affects performance of the estimator. The estimator based on cubic spline expansion performs much better in terms of MSE and coverage probability than the linear specification. Although this Monte Carlo study is small-scale, it indicates usefulness of including additional control variables to reliably estimate group-level partial effects.

6 Extensions

6.1 Instrumental Variables for Omitted Variable Bias

In the previous sections, I assume away omitted variable bias by imposing exogeneity of χ_{sg}. If instrument variables Z_{sg} for χ_{sg} are available, combination of the control function and IV methods identifies the partial effects of group-level covariates under the presence of both selection and omitted variable biases.

Assumption 9. *(i) In addition to* $\{(Y_{ig}, X_{ig}', W_{ig}', J_{ig})'\}_{1\le i\le N, 1\le g\le G}$*, a researcher observes* $\{Z_{sg}\}_{s\in\mathcal{S}, 1\le g\le G}$ *and, after redefining* $\mathbf{B}_g$ *by including* $\mathbf{Z}_g = \{Z_{sg} : s \in \mathcal{S}\}$*, Assumption 1 holds.*

(ii) For all $s \in \mathcal{S}$*,* $\mathbb{E}[\varepsilon_{isg}|F_{W|J,\mathbf{B}}(\cdot|s, \mathbf{B}_g), J_{ig} = s] = 0$ *and* $\mathbb{E}[\tilde{Z}_{ig}\varepsilon_{isg}|J_{ig} = s] = 0$ *where* $\tilde{Z}_{ig} = (Z_{ig}', W_{ig}')'$*.*

(iii) Recall $V_{ig} = (X_{ig}', W_{ig}')'$*. The matrix* $\mathbb{E}[\tilde{Z}_{ig}\{V_{ig} - \mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)]\}']$ *is of full column rank.*

Theorem 4. *Under the hypothesis of Lemma 1 and Assumption 9, β in (2) is identified.*

This result is a straightforward extension of Theorem 1 but can have an important application. For instance, if a policy intervention creates an exogenous variation in teacher assignment to different schools and households select schools after the intervention, then omitted variable bias (i.e., correlation between X_{sg} and χ_{sg}, where X_{sg} measures school-level teacher quality) can be resolved through IV but selection bias remains problematic for identification of coefficients on X_{sg}. Theorem 4 establishes that the combination of IV and control function methods achieves identification of group-level partial effects.

6.2 Nonparametric Identification

I extend the identification result to a more general outcome equation

$$Y_{isg} = m(X_{sg}, W_{ig}, \varepsilon_{isg})$$
$$J_{ig} = J(\mathbf{B}_{sg}, \Theta_{ig}, \boldsymbol{\eta}_{ig}).$$

In the linear model, I distinguish among $(\omega_{ig}, \chi_{sg}, \epsilon_{isg})$ but in this model ε_{isg} subsumes all the unobservable components due to its nonseparability. In this model, a family of parameters can be defined as

$$M(x) := \int m(x, w, e) f_{W\varepsilon}(w, e) d(w, e) \qquad x \in \mathcal{X}$$

where $f_{W\varepsilon}$ is the joint density of $(W_{ig}, \varepsilon_{isg})$ and $\mathcal{X} \subset \text{supp}(X_{sg})$ is some non-empty set. This object is called Average Structural Function (ASF) in the literature and it summarizes partial effects of a covariate X on the outcome Y. For identification, I impose the following conditions.

Assumption 10. *(i)* $(W_{ig}, \boldsymbol{\varepsilon}_{isg}) \perp\!\!\!\perp \boldsymbol{\eta}_{ig} | \Theta_{ig}, \mathbf{B}_g$ *and* $(W_{ig}, \boldsymbol{\varepsilon}_{isg}, \Theta_{ig}) \perp\!\!\!\perp \mathbf{B}_g$. *Also,* $(W_{ig}, \varepsilon_{isg})$ *has identical distributions across* $s \in \mathcal{S}$.

(ii) $\sup_{s \in \mathcal{S}, \mathbf{b} \in \text{supp}(\mathbf{B}_g)} \mathbb{E}[|m(x, W_{ig}, \varepsilon_{isg})| | J_{ig} = s, \mathbf{B}_g = \mathbf{b}] < \infty$.

(iii) Given non-empty $\mathcal{X} \subset \text{supp}(X_{sg})$, $\text{supp}(f_{W|J,\mathbf{B}}(\cdot | J_{ig}, \mathbf{B}_g))$ *is invariant conditional on* $X_{ig} = x$ *for* $x \in \mathcal{X}$.

Assumption (i) imposes independence between χ_{sg} and X_{sg}, stronger than mean exogeneity of χ_{sg} given X_{sg}. Thus, I assume away omitted variable bias and focus on the issue of selection bias. Imposition of identical distributions across $s \in \mathcal{S}$ simplifies some arguments and interpretation of the parameter. Assumption (ii) is used to justify interchanging orders of certain integrations. The requirement (iii), which states that the support of $f_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)$ is invariant conditional on X_{ig}, is referred to as support condition in the literature on triangular models, and it is considered to be a stringent assumption. Still, it holds for some special case. If the conditional support of $\mathbf{A}$ given $X_s = x$ does not vary with x, then the support condition holds. This sufficient condition excludes, among other things, that X_s is part of A_s. This may not be so restrictive if individuals only observe a coarse version of X_s when they make group decision. That is, A_s is a noisy measure of X_s.

The following theorem states the identification result for the nonseparable model.

Theorem 5. *If Assumptions 1, 2, 3, and 10 hold, $M(x)$ is identified for $x \in \mathcal{X}$.*

If the support condition is not satisfied, we can still identify different versions of ASF as done in the literature. For instance, a conditional version of ASF is

$$\int m(x, w, e) f_{W\varepsilon|f_{W|J,\mathbf{B}}}(w, e|f) d(w, e) \qquad x \in \mathcal{X}$$

where I condition on the random function $f_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)$. We can see this parameter as a measure of the partial effect conditioning on a within-group distribution of individual characteristics. For example, if ε_{isg} represents an unobservable measure of student's motivation, then the conditional ASF represents the average outcome for different levels of group-level covariate X, fixing the within-school distribution of student's motivation.

7 Conclusion

This paper presents a new identification result for group-level causal effects in a setting where individuals select into groups partially based on their unobserved heterogeneity. I build an econometric model that extends Heckman's selection model to feature group-level variables and show that group-level covariates correlate with the group-mean of the individual unobserved heterogeneity. As an alternative to instrumental variables, I exploit observable shifters to construct a valid

control function and develop a formal identification result of group-level ceteris paribus effects in a partially linear model. I propose a simple two-step semiparametric regression-based estimator, prove its consistency and asymptotic normality, and provide a consistent variance estimator. Simulation studies indicate good finite-sample properties of the proposed estimator. I also consider two extensions of the control function method. First, I combine the control function method with IVs to address another source of endogeneity, which I call omitted variable bias, and, second, I develop a nonseparable version of the model to identify the average structural functions. Finally, I empirically study the effects of school/neighborhood characteristics on student outcomes following the work of Altonji and Mansfield (2018) and find that their linear specification is robust to inclusion of additional controls.

8 Bibliography

Aaronson, D. (1998): "Using Sibling Data to Estimate the Impact of Neighborhoods on Children's Educational Outcomes," *Journal of Human Resources*, 33, 915–946.

Abdulkadiroğlu, A., P. A. Pathak, J. Schellenberg, and C. R. Walters (2017): "Do Parents Value School Effectiveness?" Working Paper.

Abdulkadiroğlu, A., P. A. Pathak, and C. R. Walters (2018): "Free to Choose: Can School Choice Reduce Student Achievement?" *American Economic Journal: Applied Economics*, 10, 175–206.

Abowd, J. M., K. L. McKinney, and I. M. Schmutte (2018): "Modeling Endogenous Mobility in Earnings Determination," Forthcoming in Journal of Business and Economic Statistics.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "Selection on Observed and Unobserved Variables: Assesing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151–184.

Altonji, J. G. and R. K. Mansfield (2018): "Estimating Group Effects Using Averages of Observables to Control for Sorting on Unobservables: School and Neighborhood Effects," *American Economic Review*, 108, 2902–2946.

Andrews, D. W. K. (2017): "Examples of L^2-Complete and Boundedly-Complete Distributions," *Journal of Econometrics*, 199, 213–220.

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

Angrist, J. D., P. A. Pathak, and C. R. Walters (2013): "Explaining Charter School Effectiveness," *American Economic Journal: Applied Economics*, 5, 1–27.

Arellano, M., R. Blundell, and S. Bonhomme (2017): "Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework," *Econometrica*, 85, 693–734.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186, 345–366.

Berry, S. T. and P. A. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797.

Blume, L. E., W. A. Brock, S. N. Durlauf, and R. Jayaraman (2015): "Linear Social Interaction Models," *Journal of Political Economy*, 123.

Blundell, R. W. and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Reviews of Economic Studies*, 71, 655–679.

Bonhomme, S., T. Lamadon, and E. Manresa (2018): "A Distributional Framework for Matched Employer Employee Data," Working paper.

Briesch, R., P. K. Chintagunta, and R. L. Matzkin (2010): "Nonparametric Discrete Choice Models with Unobserved Heterogeneity," *Journal of Business and Economic Statistics*, 28, 291–307.

Card, D. and J. Rothstein (2007): "Racial Segregation and the Black-White Test Score Gap," *Journal of Public Economics*, 91, 2158–2184.

Cattaneo, M. D., M. H. Farrell, and Y. Feng (2018): "Large Sample Properties of Partitioning-Based Series Estimators," Working Paper.

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441.

Chetty, R. and N. Hendren (2018): "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects," *Quarterly Journal of Economics*, 133, 1107–1162.

Chetty, R., N. Hendren, and L. Katz (2016): "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment," *American Economic Review*, 106, 855–902.

Chetverikov, D., B. Larsen, and C. Palmer (2016): "IV Quantile Regression for Group-Level Treatments, With an Application to the Distributional Effects of Trade," *Econometrica*, 84, 809–833.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78, 883–931.

Dahl, G. B. (2002): "Mobility and the Return to Education: Testing a Roy Model with Multiple Markets," *Econometrica*, 70, 2367–2420.

Dale, S. B. and A. B. Krueger (2002): "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables," *Quarterly Journal of Economics*, 117, 1491–1527.

Darolles, S., Y. Fan, J. P. Florens, and E. Renault (2011): "Nonparametric Instrumental Regression," *Econometrica*, 79, 1541–1565.

Das, M., W. K. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

de Paula, A. (2017): "Econometrics of Network Models," in *Advances in Economics and Econometrics:Theory and Applications, Eleventh World Congress*, ed. by B. Honore, A. Pakes, M. Piazzesi, and L. Samuelson, Cambridge University Press, 268–323.

D'Haultfoeuille, X. (2011): "On the Completeness Condition in Nonparametric Instrumental Problems," *Econometric Theory*, 27, 460–471.

Dobbie, W. and R. G. Fryer, Jr. (2011): "Are High-Quality Schools Enough to Increase Achievement Among the Poor? Evidence from the Harlem Children's Zone," *American Economic Journal:Applied Economics*, 3, 158–187.

Dubin, J. A. and D. L. McFadden (1984): "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption," *Econometrica*, 52, 345–362.

Durlauf, S. N. (1996): "Associational Redistribution: A Defense," *Politics & Society*, 24, 391–410.

——— (2004): "Neighborhood Effects," in *Handbook of Regional and Urban Economics*, ed. by J. V. Henderson and J.-F. Thisse, Elsevier, vol. 4, 2173–2242.

Durlauf, S. N. and Y. M. Ioannides (2010): "Social Interactions," *Annual Review of Economics*, 2, 451–478.

Freyberger, J. (2018): "Nonparametric Panel Data Models with Interactive Fixed Effects," *Review of Economic Studies*, 85, 1824–1851.

Gould, E. D., V. Lavy, and M. D. Paserman (2004): "Immigrating to Opportunity: Estimating the Effect of School Quality using a Natural Experiment on Ethiopians in Israel," *Quarterly Journal of Economics*, 119, 489–526.

Graham, B. S. (2017): "An Econometric Model of Network Formation with Degree Heterogeneity," *Econometrica*, 85, 1033–1063.

——— (2018): "Identifying and Estimating Neighborhood Effects," *Journal of Economic Literature*, 56, 450–500.

Hall, P. and J. L. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics*, 33, 1–27.

Hanushek, E. A., J. F. Kain, and S. G. Rivkin (2009): "New Evidence about Brown v. Board of Education: The Complex Effects of School Racial Composition on Achievement," *Journal of Labor Economics*, 27, 349–383.

Heckman, J. J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679–693.

——— (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables," *Annals of Economic and Social Measurement*, 5, 475–492.

——— (1979): "Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

Hoxby, C. M. (2000): "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, 115, 1239–1285.

Hu, Y. (2017): "The Econometrics of Unobservables: Applications of Measurement Error Models in Empirical Industrial Organization and Labor Economics," *Journal of Econometrics*, 200, 154–168.

Hu, Y. and S. M. Schennach (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76, 195–216.

Hu, Y., S. M. Schennach, and J.-L. Shiu (2017): "Injectivity of a Class of Integral Operators with Compactly Supported Kernels," *Journal of Econometrics*, 200, 48–58.

Hu, Y. and J.-L. Shiu (2018): "Nonparametric Identification Using Instrumental Variables: Sufficient Conditions for Completeness," *Econometric Theory*, 34, 659–693.

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations models without Additivity," *Econometrica*, 77, 1481–1512.

Kallenberg, O. (2017): *Random Measures, Theory and Applications*, Cham, Switzerland: Springer.

Kling, J. R., J. B. Liebman, and L. F. Katz (2007): "Experimental Analysis of Neighborhood Effects," *Econometrica*, 75, 83–119.

Kress, R. (2014): *Linear Integral Equations*, New York, NY: Springer.

Lee, L.-F. (1983): "Generalized Econometric Models with Selectivity," *Econometrica*, 51, 507–512.

Ludwig, J., G. J. Duncan, and P. Hirschfield (2001): "Urban Poverty and Juvenile Crime: Evidence from a Randomized Housing-Mobility Experiment," *Quarterly Journal of Economics*, 116, 655–679.

Mattner, L. (1993): "Some Incomplete But Boundedly Complete Location Families," *Annals of Statistics*, 21, 2158–2162.

Matzkin, R. L. (2016): "On Independence Conditions in Nonseparable Models: Observable and Unobservable Instruments," *Journal of Econometrics*, 191, 302–311.

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

Oreopoulos, P. (2003): "The Long-Run Consequences of Living in a Poor Neighborhood," *Quarterly Journal of Economics*, 118, 1533–1575.

Pollard, D. (1989): "Asymptotics via Empirical Processes," *Statistical Science*, 4, 341–366.

Sasaki, Y. (2015): "Heterogeneity and Selection in Dynamic Panel Data," *Journal of Econometrics*, 188, 236–249.

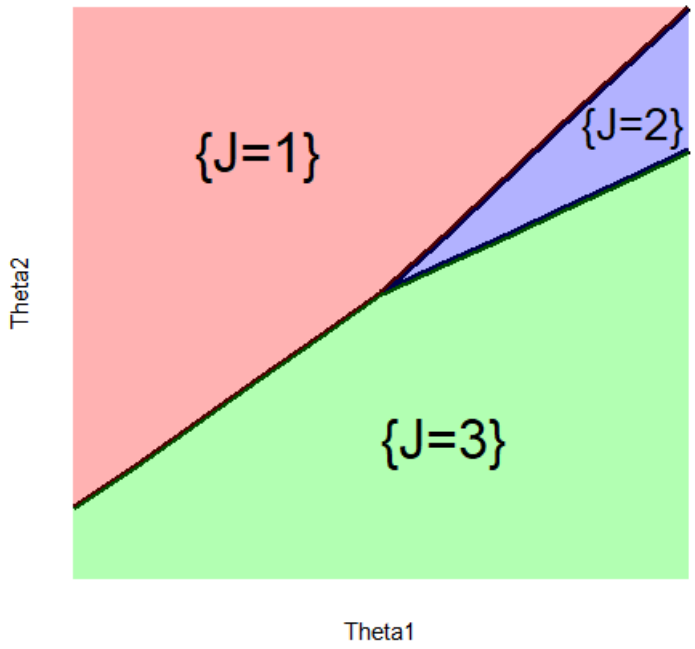
Schmitt, B. A. (1992): "Pertubation Bounds for Matrix Square Roots and Pythagorean Sums," *Linear Algebra and Its Applications*, 174, 215–227.

Sloan, F. A., G. A. Picone, D. H. Taylor, Jr., and S.-Y. Chou (2001): "Hospital Ownership and Cost and Quality of Care: Is There a Dime's Worth of Difference?" *Journal of Health Economics*, 20, 1–21.

Tropp, J. A. (2012): "User-Friendly Tail Bounds for Sums of Random Matrices," *Foundations of Computational Mathematics*, 12, 389–434.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, New York, NY: Springer.

Figure 1: Support of Θ Partitioned by Group Choice



Notes. This is an example of partitioning of the support of Θ_i by $\{J_i = 1\}$, $\{J_i = 2\}$, and $\{J_i = 3\}$ ($S = 3$). The above figure is created based on $A_1 = (1, 3)$, $A_2 = (3, 2)$, and $A_3 = (4, 1)$ with $J_i = \arg\max_{s \in \{1,2,3\}} \{\Theta_i' A_s\}$. If Θ_i falls onto the red region, $J_i = 1$ will be chosen. Similarly, the blue region represents $J_i = 2$, and the green region means $J_i = 3$.

Table 1: Estimates of School/Neighborhood Effects on Wage and Years of Post Secondary Education

	Wage w/ PS. Ed.		**Wage w/o PS. Ed.**		**Yrs. PS. Ed.**	
	AM	This Paper	AM	This Paper	AM	This Paper
Q10-90						
	0.157	0.162	0.153	0.158	0.352	0.369
	(0.017)	(0.017)	(0.017)	(0.017)	(0.079)	(0.089)
Q10-50						
	0.078	0.081	0.076	0.079	0.176	0.185
	(0.009)	(0.008)	(0.009)	(0.008)	(0.039)	(0.045)
Mean						
	2.878		2.878		1.836	

Notes. The table shows various estimates for effects of school/neighborhood quality on early adulthood wage and years of post-secondary education. The first and second row groups ("Q10-90" and "Q10-50") correspond to the impact of moving from the 10th percentile school/neighborhood to the 90th/50th percentile, respectively. The top headers indicate outcome variables in regression. "Wage w/ PS. Ed." and "Wage w/o PS. Ed." refer to log wage with/without post-secondary education as a control, respectively, and "Yrs. PS. ED." denotes years of post-secondary education. "AM" refers to the specification used in Altonji and Mansfield (2018) and "This Paper" specification adds within-school means of interaction and squared terms of W_i variables. For each row group, the number in the first line is the estimate and the number in parentheses represents the standard error estimate. The last row is the average of dependent variables in the data sample.

Table 2: Results of Monte Carlo Experiments

	Specification		
	Oracle	AM	This Paper
DGP 1			
MSE	1.00	1.69	1.75
Coverage	0.956	0.914	0.913
DGP 2			
MSE	1.00	3.88	2.36
Coverage	0.942	0.876	0.951
DGP 3			
MSE	1.00	23.32	4.45
Coverage	0.959	0.498	0.922

Notes. The table presents the results of simulation studies, particularly the mean squared error (MSE) and coverage probability of the confidence interval for different specifications. The columns "Oracle" represents infeasible regression with unobserved heterogeneity ω_i, the column "AM" corresponds to the specification of Altonji and Mansfield, and the column "This Paper" denotes the specification with group means of individual variables using cubic polynomial splines. Each row group presents different data generating processes, the row "MSE" presents (relative) MSE of the corresponding estimator divided by the MSE of the "oracle" specification, and the row "Coverage" is the coverage probability of the 95% confidence interval constructed from the corresponding estimator. The sample size is $G = 1,000$ and for each city, $N = 300$ and $S = 3$ with each group's size constraint being no more than 115 people.

Table 3: Empirical Application: Random Effects Model

	Yrs. PS. Ed.		Wage w/o PS. Ed.		Wage w/ PS. Ed.	
	AM	This Paper	AM	This Paper	AM	This Paper
Q10-90						
No Unobs.	0.215	0.267	0.121	0.123	0.125	0.122
w/ Unobs.	0.503	0.516	0.203	0.214	0.203	0.207
Q10-50						
No Unobs.	0.107	0.134	0.061	0.061	0.063	0.061
w/ Unobs.	0.251	0.258	0.102	0.107	0.102	0.104
Sample Means						
	1.620		2.880		2.880	

Notes. The estimates come from the random effects model assuming normality of the error terms. The column header indicates dependent variable used and "Q10-90" and "Q10-50" refer to effects of shifting from the 10th percentile neighborhood to the 90th/50th percentile neighborhood. The estimates are based on the lower bound on variance as described in Section 5.1. "No Unobs" rows display variance lower bound estimates excluding the random effect estimate from school/neighborhood contribution. "w/ Unobs" includes the random effect estimate as part of school/neighborhood contribution.

A Appendix: Proofs

Let $L^p(\mu)$ be the class of functions whose pth power is integrable with respect to μ. I write F_W, F_Θ for the distribution of random vectors W, Θ, and so on.

A.1 Proof of Lemma 1

Inverse of $\mathbf{\Psi}$

Recall

$$\mathbf{\Psi} h(w) = \int h(\theta) f_{\Theta|W}(\theta|w) d\lambda(\theta).$$

Let $\mathcal{H}$ be the range of $\mathbf{\Psi}$ on $L^2(F_\Theta)$. I show that $h \in \mathcal{H}$ is square-integrable with respect to F_W. For any $h \in L^2(F_\Theta)$,

$$\begin{aligned}
\int (\mathbf{\Psi} h)^2 dF_W &= \int \Big[\int h(\theta) f_{\Theta|W}(\theta|w) d\lambda(\theta) \Big]^2 dF_W(w) \\
&= \int \big(\mathbb{E}[h(\Theta_{ig})|W_{ig} = w] \big)^2 dF_W(w) \\
&\leq \int \big[\mathbb{E}[|h(\Theta_{ig})|^2 | W_{ig} = w] dF_W(w) \\
&= \mathbb{E}[|h(\Theta_{ig})|^2] < \infty.
\end{aligned}$$

By boundedness of $f_{\Theta|W}$, $\mathbf{\Psi}$ is a compact linear operator. Then, Theorem 15.16 in Kress (2014) implies there exist non-negative reals $\{\tau_j : j \in \mathbb{N}\}$ and orthonormal sequences $\{\phi_j : j \in \mathbb{N}\}$, $\{\varphi_j : j \in \mathbb{N}\}$ in $L^2(F_\Theta)$ and $\mathcal{H}$, respectively, such that

$$\mathbf{\Psi} h = \sum_{j=1}^{\infty} \tau_j \langle h, \phi_j \rangle_\Theta \varphi_j$$

where $\langle h, \phi \rangle_\Theta = \int h \phi dF_\Theta$. Also let $\langle m, \varphi \rangle_W = \int m \varphi dF_W$.

Without loss of generality, assume $\tau_j > 0$ for all j. Define

$$\mathbf{\Psi}^\dagger m = \sum_{j=1}^{\infty} \tau_j^{-1} \langle m, \varphi_j \rangle_W \phi_j.$$

By definition, $\mathbf{\Psi}^\dagger \mathbf{\Psi} h = h$ on $L^2(F_\Theta)$.

Integral with respect to $f_{W|J,\mathbf{B}}$

$f_{W|J,\mathbf{B}}/f_W \in \mathcal{H}$ because

$$\begin{aligned} f_{W|J,\mathbf{B}}(w|s,\mathbf{b}) &= \int f_{W|\Theta,J,\mathbf{B}}(w|\theta,s,\mathbf{b}) f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\lambda(\theta) \\ &= \int f_{W|\Theta,\mathbf{B}}(w|\theta,\mathbf{b}) f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\lambda(\theta) \\ &= \int f_{W|\Theta}(w|\theta) f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\lambda(\theta) \\ &= \mathbf{\Psi}(f_{\Theta|J,\mathbf{B}}/f_\Theta)(w) f_W(w) \end{aligned}$$

and $f_{\Theta|J,\mathbf{B}}/f_\Theta \in L^2(F_\Theta)$, where the second equality holds by $J = J(\Theta, \mathbf{A}, \boldsymbol{\eta})$ and $\boldsymbol{\eta} \perp\!\!\!\perp (W, \Theta)|\mathbf{B}$ and the third follows from independence $W \perp\!\!\!\perp \mathbf{B}|\Theta$.

Then, $\mathbf{\Psi}^\dagger(f_{W|J,\mathbf{B}}/f_W) = f_{\Theta|J,\mathbf{B}}/f_\Theta$. Write $m(\theta) = \mathbb{E}[\omega_{ig}|\Theta_{ig} = \theta]$. If the interchange of integral and infinite sum is permitted,

$$\begin{aligned} \mathbb{E}[\omega_{ig}|J_{ig} = s, \mathbf{B}_g = \mathbf{b}] &= \int \mathbb{E}[\omega_{ig}|\Theta_{ig} = \theta, J_{ig} = s, \mathbf{B}_g = \mathbf{b}] f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\lambda(\theta) \\ &= \int \mathbb{E}[\omega_{ig}|\Theta_{ig} = \theta] f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\lambda(\theta) \\ &= \int m(\theta) \frac{f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b})}{f_\Theta(\theta)} dF_\Theta(\theta) \\ &= \int m(\theta) \big(\mathbf{\Psi}^\dagger f_{W|J,\mathbf{B}}(\cdot|s,\mathbf{b})/f_W(\cdot)\big)(\theta) dF_\Theta(\theta) \\ &= \int m(\theta) \left[\sum_{j=1}^\infty \tau_j^{-1} \langle f_{W|J,\mathbf{B}}(\cdot|s,\mathbf{b})/f_W(\cdot), \varphi_j\rangle_W \phi_j(\theta)\right] dF_\Theta(\theta) \\ &= \sum_{j=1}^\infty \tau_j^{-1} \langle f_{W|J,\mathbf{B}}(\cdot|s,\mathbf{b})/f_W(\cdot), \varphi_j\rangle_W \int m(\theta)\phi_j(\theta) dF_\Theta(\theta) \qquad (12) \\ &= \int \left[\sum_{j=1}^\infty \tau_j^{-1} \int m(\theta)\phi_j(\theta) dF_\Theta(\theta) \varphi_j(w)\right] f_{W|J,\mathbf{B}}(w|s,\mathbf{b}) d\mu(w) \qquad (13) \end{aligned}$$

where I use $\omega_{ig} \perp\!\!\!\perp \mathbf{B}_g|\Theta_{ig}$ for the second equality. To justify (12), it suffices to show

$$\sum_{j=1}^\infty \int \tau_j^{-1} \big|\langle f_{W|J,\mathbf{B}}(\cdot|s,\mathbf{b})/f_W(\cdot), \varphi_j\rangle_W\big| |m(\theta)\phi_j(\theta)| dF_\Theta(\theta) < \infty.$$

Using $f_{W|J,\mathbf{B}}/f_W = \mathbf{\Psi} f_{\Theta|J,\mathbf{B}}/f_\Theta$,

$$\begin{aligned}\langle f_{W|J,\mathbf{B}}(\cdot|s,\mathbf{b})/f_W(\cdot), \varphi_j\rangle_W &= \langle \mathbf{\Psi}(f_{\Theta|J,\mathbf{B}}/f_\Theta), \varphi_j\rangle_W \\ &= \Big\langle \sum_{\ell=1}^{\infty} \lambda_\ell \langle f_{\Theta|J,\mathbf{B}}/f_\Theta, \phi_\ell\rangle_\Theta \varphi_\ell, \varphi_j \Big\rangle_W \\ &= \sum_{\ell=1}^{\infty} \lambda_\ell \langle f_{\Theta|J,\mathbf{B}}/f_\Theta, \phi_\ell\rangle_\Theta \langle \varphi_\ell, \varphi_j\rangle_W \\ &= \tau_j \langle f_{\Theta|J,\mathbf{B}}/f_\Theta, \phi_j\rangle_\Theta.\end{aligned}$$

Thus, it suffices to have

$$\sum_{j=1}^{\infty} \big|\langle f_{\Theta|J,\mathbf{B}}/f_\Theta, \phi_j\rangle_\Theta\big| \langle |m|, |\phi_j|\rangle_\Theta \le \|m\|_\Theta \sum_{j=1}^{\infty} \big|\langle f_{\Theta|J,\mathbf{B}}/f_\Theta, \phi_j\rangle_\Theta\big| < \infty$$

where $\|\cdot\|_\Theta$ is the norm induced by $\langle\cdot,\cdot\rangle_\Theta$ and I use $\|\phi_j\|_\Theta = 1$ for $j \in \mathbb{N}$. For (13), it suffices to show

$$\begin{aligned}&\sum_{j=1}^{\infty} \int \tau_j^{-1} |f_{W|J,\mathbf{B}}(w|s,\mathbf{b})\varphi_j(w)| \big|\langle m, \phi_j\rangle_\Theta\big| d\mu(w) \\ &\le \sum_{j=1}^{\infty} \tau_j^{-1} \big|\langle m, \phi_j\rangle_\Theta\big| \Big[\int |f_{W|J,\mathbf{B}}(w|s,\mathbf{b})|^2 d\mu(w) \int |\varphi_j(w)|^2 d\mu(w) \Big]^{1/2} \\ &\le C \sum_{j=1}^{\infty} \tau_j^{-1} \big|\langle m, \phi_j\rangle_\Theta\big| < \infty.\end{aligned}$$

Thus, the interchageability follows from Assumption 4.

Since

$$\psi(w) = \sum_{j=1}^{\infty} \tau_j^{-1} \alpha_j \varphi_j(w), \qquad \alpha_j = \langle m, \phi_j\rangle_\Theta,$$

$\{\alpha_j/\tau_j : j \in \mathbb{N}\}$ is square-summable, and $\{\varphi_j : j \in \mathbb{N}\}$ is an orthonormal set with respect to F_W, ψ is square integrable with respect to F_W.

A.2 Proof of Theorem 1

By taking conditional expectations given $F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)$,

$$\begin{aligned}&\mathbb{E}[Y_{iJ_{ig}g}|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]\\&=\mathbb{E}[X_{J_{ig}g}|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]'\beta+\mathbb{E}[W_{ig}|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]'\gamma+\int\psi(w)dF_{W|J,\mathbf{B}}(w|J_{ig},\mathbf{B}_g)\end{aligned}$$

and subtracting this conditional expectation from the equation (11),

$$\begin{aligned}&Y_{iJ_{ig}g}-\mathbb{E}[Y_{iJ_{ig}g}|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]\\&=\left\{X_{J_{ig}g}-\mathbb{E}[X_{J_{ig}g}|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]\right\}'\beta+\left\{W_{ig}-E[W_{ig}|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]\right\}'\gamma+\varepsilon_{ig}.\end{aligned}$$

Writing $V_{ig}=(X'_{J_{ig}g}\,W'_{ig})'$,

$$\mathbb{E}[\{V_{ig}-\mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}_g}]\}Y_{iJ_{ig}g}]=\mathbb{E}[\{V_{ig}-\mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}_g}]\}\{V_{ig}-\mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}_g}]\}']\begin{pmatrix}\beta\\\gamma\end{pmatrix}.$$

Then, the invertibility condition (ii) guarantees identifiability of β.

A.3 Proof of Theorem 2

Below, I use the following notational conventions. For two sequences of real numbers a_n and b_n, $a_n\lesssim b_n$ means there exists a constant C not dependent on n such that $a_n/b_n\leq C$ for n large enough. For sequences of random variables X_n and Y_n, $X_n\lesssim_{\mathbb{P}}Y_n$ means that there exists a fixed constant C satisfying $\Pr(X_n\leq CY_n)\to 1$ as $n\to\infty$. Given a symmetric matrix $\mathbf{X}$, I write $\lambda_{\max}(\mathbf{X})$ and $\lambda_{\min}(\mathbf{X})$ for largest and smallest eigenvalues of the matrix $\mathbf{X}$. Define $\bar{\lambda}_K=\lambda_{\max}(\mathbf{Q}_K)$ and $\tilde{\lambda}=1/\lambda_{\min}(\mathbf{Q}_K)$.

Also, define

$$\begin{aligned}\boldsymbol{r}_K&=\Big(\sum_{k=K+1}^{\infty}\delta_k\mathbb{E}[\rho_k(W_{11})|J_{11},\mathbf{B}_1],\ldots,\sum_{k=K+1}^{\infty}\delta_k\mathbb{E}[\rho_k(W_{NG})|J_{NG},\mathbf{B}_G]\Big)'\\\boldsymbol{\varepsilon}&=(\varepsilon_{11},\ldots,\varepsilon_{NG})'.\end{aligned}$$

Proof. Lemma 2 below implies the first part. By $\mathrm{Var}(\chi_s-\mathbb{E}[\omega_{ig}|J_{ig}=s,\mathbf{B}_g])>0$, $\mathbb{E}[\boldsymbol{\varepsilon}_{isg}\boldsymbol{\varepsilon}_{jsg}]>0$

and the minimum eigenvalue of

$$\mathbb{E}[\psi_{Ng}\psi_{Ng}'] = \frac{1}{N}\mathbb{E}[P_{Kig}P_{Kig}'\varepsilon_{ig}^2] + (1-1/N)\mathbb{E}[P_{Kig}P_{Kjg}'\varepsilon_{ig}\varepsilon_{jg}]$$

is bounded from below uniformly in K. In addition, $\|\mathbf{Q}_K^{-1}\| \geq 1/\lambda_{\max}(\mathbf{Q}_K)$, which is bounded away from zero uniformly. Thus, $\lambda_{\min}(\mathbf{\Omega}_K)$ is bounded away from zero uniformly.

For the asymptotic normality of the oracle estimator, note ψ_{Ng} is mean zero and independent across t. I verify Lindberg's condition. The variance matrix is the identity matrix by construction. For any vector $\mathbf{v}$, I want to show for all $d > 0$,

$$\mathbb{E}[|\mathbf{v}'\mathbf{\Omega}_K^{-1/2}\mathbf{S}\mathbf{Q}_K^{-1}\psi_{Ng}|^2\mathbb{1}\{|\mathbf{v}'\mathbf{\Omega}_K^{-1/2}\mathbf{S}\mathbf{Q}_K^{-1}\psi_{Ng}| \geq d\sqrt{G}\}] \to 0.$$

Letting $\varpi_{ig} = \mathbf{v}'\mathbf{\Omega}_K^{-1/2}\mathbf{S}\mathbf{Q}_K^{-1}P_{Kig}$, we have $\mathbf{v}'\mathbf{\Omega}_K^{-1/2}\mathbf{S}\mathbf{Q}_K^{-1}\psi_{Ng} = \sum_i \varpi_{ig}\varepsilon_{ig}/N$ and $|\varpi_{ig}| \lesssim \zeta_K$.

$$\begin{aligned}
&\mathbb{E}[|\mathbf{v}'\mathbf{\Omega}_K^{-1/2}\mathbf{S}\mathbf{Q}_K^{-1}\psi_{Ng}|^2\mathbb{1}\{|\mathbf{v}'\mathbf{\Omega}_K^{-1/2}\mathbf{S}\mathbf{Q}_K^{-1}\psi_{Ng}| \geq d\sqrt{G}\}] \\
&\leq \mathbb{E}\frac{1}{N}\sum_{i=1}^N |\varpi_{ig}|^2 \frac{1}{N}\sum_{i=1}^N |\varepsilon_{ig}|^2 \mathbb{1}\Big\{\Big|\frac{1}{N}\sum_{i=1}^N |\varepsilon_{ig}|^2\Big|^{1/2} \geq d\sqrt{G}/\zeta_K\Big\} \\
&\leq \mathbb{E}[|\varpi_{ig}|^2] \sup_{\mathbf{P}_{Kg}} \mathbb{E}\Big[\frac{1}{N}\sum_{i=1}^N |\varepsilon_{ig}|^2 \mathbb{1}\Big\{\Big|\frac{1}{N}\sum_{i=1}^N |\varepsilon_{ig}|^2\Big|^{1/2} \geq d\sqrt{G}/\zeta_K\Big\}\Big|\mathbf{P}_{Kg}\Big] \\
&\leq \mathbb{E}[|\varpi_{ig}|^2]\frac{\zeta_K^2}{d^2G} \sup_{\mathbf{P}_{Kg}} \mathbb{E}\Big[\Big(\frac{1}{N}\sum_{i=1}^N |\varepsilon_{ig}|^2\Big)^2\Big|\mathbf{P}_{Kg}\Big] \lesssim \zeta_K^4/G \to 0.
\end{aligned}$$

□

Now I prove the asymptotic (first-order) equivalency between the oracle and feasible estimators. Let $\theta = (\beta', \gamma', \delta_1, \ldots, \delta_K)$ and we have

$$\begin{aligned}
\hat{\beta} - \beta &= \mathbf{S}\tilde{\mathbf{Q}}_K^{-1}\hat{\mathbf{P}}_K'(\mathbf{P}_K - \hat{\mathbf{P}}_K)\theta/NG + \mathbf{S}\tilde{\mathbf{Q}}_K^{-1}\hat{\mathbf{P}}_K'\boldsymbol{r}_K/NG + \mathbf{S}\tilde{\mathbf{Q}}_K^{-1}\hat{\mathbf{P}}_K'\boldsymbol{\varepsilon}/NG \\
&\equiv \mathrm{I} + \mathrm{II} + \mathrm{III}.
\end{aligned}$$

For I, Lemmas 2, 3, and 5 imply that it is $O_{\mathbb{P}}(\tilde{\theta}_K \max\{\zeta_K, \sqrt{K}\}\sqrt{K\log(NG)/NG})$. For II, the

calculation done in Lemma 2 yields that it is $O_{\mathbb{P}}(\sqrt{\tilde{\lambda}_K}K^{-b})$, and for III,

$$\begin{aligned}\mathbf{S}\tilde{\mathbf{Q}}_K^{-1}\hat{\mathbf{P}}'_K\boldsymbol{\varepsilon}/NG &= \mathbf{S}\{\tilde{\mathbf{Q}}_K^{-1}-\hat{\mathbf{Q}}_K^{-1}\}(\hat{\mathbf{P}}_K-\mathbf{P}_K)'\boldsymbol{\varepsilon}/NG+\mathbf{S}\{\tilde{\mathbf{Q}}_K^{-1}-\hat{\mathbf{Q}}_K^{-1}\}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG\\ &\quad+\hat{\mathbf{Q}}_K^{-1}(\hat{\mathbf{P}}_K-\mathbf{P}_K)'\boldsymbol{\varepsilon}/NG+\hat{\mathbf{Q}}_K^{-1}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG\\ &= \hat{\mathbf{Q}}_K^{-1}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG+O_{\mathbb{P}}(\tilde{\lambda}_K^2\max\{\zeta_K,\sqrt{K}\}\sqrt{K\log(NG)/NG}+\tilde{\lambda}_K\sqrt{K/NG}).\end{aligned}$$

A.3.1 Lemmas for Proof of Theorem 2

Lemma 2. *Assume* $\zeta_K^2\tilde{\lambda}_K^2\bar{\lambda}_K\log K/G\to 0$.

$$\hat{\beta}_{oracle}-\beta=\mathbf{S}\mathbf{Q}_K^{-1}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG+O_{\mathbb{P}}((\tilde{\lambda}_K)^{1/2}K^{-b})+O_{\mathbb{P}}\Big(\frac{\tilde{\lambda}_K^2\zeta_K\bar{\lambda}_K\sqrt{\log K}}{G}\Big)$$

where $\zeta_K=(NG)^{1/m_w}+G^{1/m_x}+\xi_K$.

Proof. By decomposition, $\hat{\beta}_{\text{oracle}}=\beta+\mathbf{S}(\mathbf{P}'_K\mathbf{P}_K)^{-1}\mathbf{P}'_K(\boldsymbol{r}_K+\boldsymbol{\varepsilon})$ and

$$\hat{\beta}_{\text{oracle}}-\beta=\mathbf{S}\mathbf{Q}_K^{-1}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG+\mathbf{S}(\mathbf{P}'_K\mathbf{P}_K)^{-1}\mathbf{P}'_K\boldsymbol{r}_K+\mathbf{S}\big\{(\mathbf{P}'_K\mathbf{P}_K/NG)^{-1}-\mathbf{Q}_K^{-1}\big\}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG.$$

First, Lemma 6.2 in Belloni et al. (2015) implies

$$\|(\mathbf{P}'_K\mathbf{P}_K/NG)-\mathbf{Q}_K\|=O_{\mathbb{P}}\Big(\sqrt{\frac{\zeta_K^2\bar{\lambda}_K\log K}{G}}\Big). \tag{14}$$

Thus, the smallest eigenvalue of $\hat{\mathbf{Q}}_K=(\mathbf{P}'_K\mathbf{P}_K/NG)$ is bounded below by $\lambda_{\min}(\mathbf{Q}_K)\{1+o_{\mathbb{P}}(1)\}$ if $\zeta_K^2\tilde{\lambda}_K^2\bar{\lambda}_K\log K/G\to 0$. Using Frobenius norm $\|\cdot\|_F$,

$$\begin{aligned}\|(\mathbf{P}'_K\mathbf{P}_K)^{-1}\mathbf{P}'_K\boldsymbol{r}_K\|_F^2 &= \boldsymbol{r}'_K\mathbf{P}_K\hat{\mathbf{Q}}_K^{-1/2}\hat{\mathbf{Q}}_K^{-1}\hat{\mathbf{Q}}_K^{-1/2}\mathbf{P}'_K\boldsymbol{r}_K/(NG)^2\\ &\lesssim_{\mathbb{P}}\tilde{\lambda}_K\boldsymbol{r}'_K\mathbf{P}_K(\mathbf{P}'_K\mathbf{P}_K)^{-1}\mathbf{P}'_K\boldsymbol{r}_K/NG\\ &\lesssim\tilde{\lambda}_K\boldsymbol{r}'_K\boldsymbol{r}_K/NG=O_{\mathbb{P}}(\tilde{\lambda}_K K^{-2b}).\end{aligned}$$

For $\mathbf{S}\{(\mathbf{P}'_K\mathbf{P}_K/NG)^{-1}-\mathbf{Q}_K^{-1}\}\mathbf{P}'_K\boldsymbol{\varepsilon}/NG$, since ε_{ig} is conditionally mean zero given P_{Kig}, it suffices

to look at its variance. For the conditional variance given $\mathbf{P}_K$,

$$N^{-2}G^{-2}\Big\|\mathbf{S}\Big\{\Big(\frac{\mathbf{P}_K'\mathbf{P}_K}{NG}\Big)^{-1}-\mathbf{Q}_K^{-1}\Big\}\sum_{g=1}^{G}\mathbf{P}_{Kg}'\mathbb{E}[\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}_g'|\mathbf{P}_{Kg}]\mathbf{P}_{Kg}\Big\{\Big(\frac{\mathbf{P}_K'\mathbf{P}_K}{NG}\Big)^{-1}-\mathbf{Q}_K^{-1}\Big\}'\mathbf{S}'\Big\|$$
$$\leq G^{-1}\big\|\mathbf{S}\big\{(\mathbf{P}_K'\mathbf{P}_K/NG)^{-1}-\mathbf{Q}_K^{-1}\big\}\frac{1}{NG}\sum_{g=1}^{G}\sum_{i=1}^{N}P_{Kig}P_{Kig}'\big\{(\mathbf{P}_K'\mathbf{P}_K/NG)^{-1}-\mathbf{Q}_K^{-1}\big\}'\mathbf{S}'\big\|$$
$$=G^{-1}\big\|\mathbf{S}\big\{(\mathbf{P}_K'\mathbf{P}_K/NG)^{-1}-\mathbf{Q}_K^{-1}\big\}\big\|^2\|\mathbf{P}_K'\mathbf{P}_K/NG\|=O_{\mathbb{P}}\Big(G^{-1}\tilde{\lambda}_K^4\frac{\zeta_K^2\bar{\lambda}_K^2\log K}{G}\Big)$$

where I use $A^{-1}-B^{-1}=B^{-1}(B-A)A^{-1}$. □

Lemma 3. *Let* $\tilde{\mathbf{Q}}_K=(\hat{\mathbf{P}}_K'\hat{\mathbf{P}}_K/NG)^{-1}$. *Under* $\max\{\zeta_K,\sqrt{K}\}\tilde{\lambda}_K\sqrt{K\log(NG)/NG}\to 0$,

$$\|\tilde{\mathbf{Q}}_K^{-1}\|\lesssim_{\mathbb{P}}\|\hat{\mathbf{Q}}_K^{-1}\|$$

Proof. Using Lemma 5

$$\|\hat{\mathbf{P}}_K'\hat{\mathbf{P}}_K/NG-\mathbf{P}_K'\mathbf{P}_K/NG\|\leq\|\hat{\mathbf{P}}_K-\mathbf{P}_K\|^2/NG+2\|(\hat{\mathbf{P}}_K-\mathbf{P}_K\|/NG$$
$$=O_{\mathbb{P}}\big(\max\{\zeta_K,\sqrt{K}\}\sqrt{K\log(NG)/NG}\big).$$

□

Lemma 4. *If Assumption 6 holds, then*

$$\max_{\substack{1\leq g\leq G\\1\leq i\leq N}}\|\hat{P}_{Kig}-P_{Kig}\|=O_{\mathbb{P}}(\sqrt{K/N}).$$

Proof. By definition,

$$\|\hat{P}_{Kig}-P_{Kig}\|^2=\sum_{\ell=1}^{K}\Big(\frac{1}{N_{J_{ig}g}}\sum_{j=1}^{N}\rho_\ell(W_{jg})\mathbb{1}\{J_{jg}=J_{ig}\}-\mathbb{E}[\rho_\ell(W_{jg})|J_{jg}=J_{ig},\mathbf{B}_g]\Big)^2$$

and using $N_{sg}/N\geq c>0$ for all $s\in\mathcal{S}$ and t with probability approaching one,

$$\frac{1}{N_{J_{ig}g}}\sum_{j=1}^{N}\rho_\ell(W_{jg})\mathbb{1}\{J_{jg}=J_{ig}\}-\mathbb{E}[\rho_\ell(W_{jg})|J_{jg}=J_{ig},\mathbf{B}_g]$$
$$=O_{\mathbb{P}}(1)\frac{1}{N}\sum_{j\neq i}\big(\rho_\ell(W_{jg})\mathbb{1}\{J_{jg}=J_{ig}\}-\mathbb{E}[\rho_\ell(W_{jg})|J_{jg}=J_{ig},\mathbf{B}_g]\big)+\frac{\rho(W_{ig})}{N}O_{\mathbb{P}}(1).$$

Conditional on J_{ig} and $\mathbf{B}_g$, the summand is independent across j and Corollary 4.3 in Pollard (1989) implies that

$$\sup_{\ell\in\mathbb{N}}\left|\frac{1}{\sqrt{N}}\sum_{j\neq i}\left(\rho_\ell(W_{jg})\mathbb{1}\{J_{jg}=J_{ig}\}-\mathbb{E}[\rho_\ell(W_{jg})|J_{jg}=J_{ig},\mathbf{B}_g]\right)\right|=O_{\mathbb{P}}(1).$$

The uniformity of the bound with respect to the conditioning variables $(J_{ig},\mathbf{B}_g)$ indicates that the above bound hold uniformly in $(i,g)\in\{1,\dots,N\}\times\{1,\dots,G\}$. □

Lemma 5. *Under the hypothesis of Lemma 4,*

$$\|\mathbf{P}_K'(\hat{\mathbf{P}}_K-\mathbf{P}_K)\|=O_{\mathbb{P}}(\sqrt{GNK\log(NG)}\max\{\zeta_K,\sqrt{K}\}).$$

Proof.

$$\mathbf{P}_K'(\hat{\mathbf{P}}_K-\mathbf{P}_K)=\sum_{g=1}^{G}\sum_{i=1}^{N}P_{Kig}(\hat{P}_{Kig}-P_{Kig})'\equiv\sum_{g=1}^{G}\Xi_g.$$

I use Theorem 1.6 in Tropp (2012). Using Lemma 4, $\|\hat{P}_{Kig}-P_{Kig}\|\leq C\sqrt{K/N}$ with probability arbitrary close to one with some large constant C and

$$\|\Xi_g\|\leq C\sum_{i=1}^{N}\sum_{j\neq i}\|P_{Kig}\|\sqrt{K/N}\leq C\zeta_K\sqrt{NK}$$

For the variance terms,

$$\begin{aligned}
\|\mathbb{E}[\Xi_g\Xi_g']\| &= \left\|\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[P_{Kig}P_{Kjg}'(\hat{P}_{Kig}-P_{Kig})'(\hat{P}_{Kjg}-P_{Kjg})]\right\| \\
&\lesssim \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[\|P_{Kig}P_{Kjg}'\|]K/N\leq N^2\zeta_K^2(K/N)=\zeta_K^2NK \\
\|\mathbb{E}[\Xi_g'\Xi_g]\| &= \left\|\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[(\hat{P}_{Kjg}-P_{Kjg})(\hat{P}_{Kig}-P_{Kig})'P_{Kig}'P_{Kjg}]\right\| \\
&\leq \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[\|\hat{P}_{Kjg}-P_{Kjg}\|\|\hat{P}_{Kig}-P_{Kig}\|\|P_{Kig}'P_{Kjg}\|]\lesssim N^2(K/N)K=K^2N.
\end{aligned}$$

Then, Theorem 1.6 in Tropp (2012) implies

$$\Pr\left(\left\|\sum_{g=1}^{G}\Xi_g\right\|\geq M\right)\leq 2NG\exp\left(\frac{-M^2/2}{NGK\max\{\zeta_K^2,K\}+\zeta_K\sqrt{NK}M/3}\right)$$

and taking $M = \sqrt{TNK\log(NG)}\max\{\zeta_K, \sqrt{K}\}$ shows the desired result. □

Lemma 6. *If Assumptions 6 and 8 hold,*

$$\|(\hat{\mathbf{P}}_K - \mathbf{P}_K)'\boldsymbol{\varepsilon}/NG\| = O_{\mathbb{P}}(\sqrt{K/NG}).$$

Proof.

$$\begin{aligned}
\mathbb{E}\|(\hat{\mathbf{P}}_K - \mathbf{P}_K)'\boldsymbol{\varepsilon}/NG\|^2 &= \mathbb{E}\Big\|\frac{1}{NG}\sum_{g=1}^{G}\sum_{i=1}^{N}(\hat{P}_{Kig} - P_{Kig})\varepsilon_{ig}\Big\|^2 \\
&= \frac{1}{N^2G^2}\sum_{g=1}^{G}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[(\hat{P}_{Kig} - P_{Kig})(\hat{P}_{Kjg} - P_{Kjg})\varepsilon_{ig}\varepsilon_{jg}] \\
&\lesssim G^{-1}\frac{K}{N}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}[|\varepsilon_{ig}\varepsilon_{jg}|].
\end{aligned}$$

□

A.4 Proof of Theorem 3

Recall

$$\hat{\boldsymbol{\Omega}}_K = \mathbf{S}\tilde{\mathbf{Q}}_K^{-1}\hat{\boldsymbol{\Sigma}}_K\tilde{\mathbf{Q}}_K^{-1}\mathbf{S}'$$

where $\tilde{\mathbf{Q}}_K = \hat{\mathbf{P}}_K'\hat{\mathbf{P}}_K/NG$,

$$\hat{\boldsymbol{\Sigma}}_K = \frac{1}{G}\sum_{g=1}^{G}\hat{\psi}_{Ng}\hat{\psi}_{Ng}' \qquad \hat{\psi}_{Ng} = \frac{1}{N}\sum_{i=1}^{N}\hat{P}_{Kig}(Y_{ig} - \hat{P}_{Kig}'\hat{\theta})$$

and $\hat{\theta} = (\hat{\mathbf{P}}_K'\hat{\mathbf{P}}_K)^{-1}(\hat{\mathbf{P}}_K'\mathbf{Y})$. Using the lemmas below,

$$\|\hat{\boldsymbol{\Omega}}_K^{-1/2} - \boldsymbol{\Omega}_K^{-1/2}\| \le \|\hat{\boldsymbol{\Omega}}_K^{-1/2}\|\|\hat{\boldsymbol{\Omega}}_K^{1/2} - \boldsymbol{\Omega}_K^{1/2}\|\|\boldsymbol{\Omega}_K^{-1/2}\| \le \tilde{\lambda}_K\{2\bar{\lambda}_K^{-1/2}\}^{-1}\|\hat{\boldsymbol{\Omega}}_K - \boldsymbol{\Omega}_K\|$$

and

$$\|\hat{\boldsymbol{\Omega}}_K - \boldsymbol{\Omega}_K\| = O_{\mathbb{P}}(\|\hat{\mathbf{Q}}_K^{-1} - \mathbf{Q}_K^{-1}\|\|\hat{\boldsymbol{\Sigma}}_K\|\|\hat{\mathbf{Q}}_K^{-1}\| + \|\hat{\mathbf{Q}}_K^{-1}\|^2\|\hat{\boldsymbol{\Sigma}}_K - \boldsymbol{\Sigma}_K\|),$$

which shows the result.

Lemma 7. *Under the hypothesis of Theorem 3,*

$$\left\|\hat{\mathbf{\Sigma}}_K - \mathbf{\Sigma}\right\| = O_{\mathbb{P}}(\zeta_K^2(\zeta_K\sqrt{K/G} + \sqrt{K/N} + K^{-\alpha}) + \zeta_K\sqrt{(NG)^{1/m_\varepsilon}\log K/G}).$$

Proof.

$$\|\hat{\mathbf{\Omega}}_K - \mathbf{\Omega}_K\| \leq \|\hat{\mathbf{\Omega}}_K - \frac{1}{G}\sum_{g=1}^{G}\psi_{Ng}\psi_{Ng}'\| + \|\frac{1}{G}\sum_{g=1}^{G}\psi_{Ng}\psi_{Ng}' - \mathbf{\Omega}_K\|.$$

For the second term, Lemma 6.2 in Belloni et al. (2015) implies that it is $O_{\mathbb{P}}(\zeta_K\sqrt{(NG)^{1/m_\varepsilon}\log K/G})$. For the first term,

$$\hat{\mathbf{\Sigma}}_K = \frac{1}{G}\sum_{g=1}^{G}(\hat{\psi}_{Ng} - \psi_{Ng})\hat{\psi}_{Ng}' + \frac{1}{G}\sum_{g=1}^{G}\psi_{Ng}(\hat{\psi}_{Ng} - \psi_{Ng})' + \frac{1}{G}\sum_{g=1}^{G}\psi_{Ng}\psi_{Ng}'$$

and

$$\begin{aligned}
\hat{\psi}_{Ng} - \psi_{Ng} &= \frac{1}{N}\sum_{i=1}^{N}(\hat{P}_{Kig} - P_{Kig})(Y_{ig} - \hat{P}_{Kig}'\hat{\theta}) - \frac{1}{N}\sum_{i=1}^{N}P_{Kig}(\hat{P}_{Kig}'\hat{\theta} - P_{Kig}'\theta) + \frac{1}{N}\sum_{i=1}^{N}P_{Kig}r_{ig} \\
&= \frac{1}{N}\sum_{i=1}^{N}(\hat{P}_{Kig} - P_{Kig})(\varepsilon_{ig} + r_{ig}) + \frac{1}{N}\sum_{i=1}^{N}(\hat{P}_{Kig} - P_{Kig})(P_{Kig}'\theta - \hat{P}_{Kig}'\hat{\theta}) \\
&\quad - \frac{1}{N}\sum_{i=1}^{N}P_{Kig}(\hat{P}_{Kig}'\hat{\theta} - P_{Kig}'\theta) + \frac{1}{N}\sum_{i=1}^{N}P_{Kig}r_{ig}.
\end{aligned}$$

Note $\|\hat{P}_{Kig} - P_{Kig}\| = O_{\mathbb{P}}(\sqrt{K/N})$ and $\|\hat{\theta} - \theta\| = O_{\mathbb{P}}(\sqrt{K/G})$. The latter follows from

$$\|\hat{\theta} - \theta\| = \|\mathbf{Q}_K^{-1}\mathbf{P}_K'\boldsymbol{\varepsilon}/NG\| + o_{\mathbb{P}}(G^{-1/2}) = O_{\mathbb{P}}(\sqrt{K/G}).$$

Then,

$$\begin{aligned}
\|P_{Kig}'\theta - \hat{P}_{Kig}'\hat{\theta}\| &\leq \|P_{Kig}\|\|\hat{\theta} - \theta\| + \|\hat{P}_{Kig} - P_{Kig}\|\hat{\theta}\| \\
&= O_{\mathbb{P}}(\zeta_K\sqrt{K/G} + \sqrt{K/N})
\end{aligned}$$

and

$$\left\| \frac{1}{N} \sum_{i=1}^{N} (\hat{P}_{Kig} - P_{Kig})(\varepsilon_{ig} + r_{ig}) \right\| = O_{\mathbb{P}}(\sqrt{K/N})$$
$$\left\| \frac{1}{N} \sum_{i=1}^{N} (\hat{P}_{Kig} - P_{Kig})(P'_{Kig}\theta - \hat{P}'_{Kig}\hat{\theta}) \right\| = O_{\mathbb{P}}(K/N + \zeta_K K/\sqrt{NG})$$
$$\left\| \frac{1}{N} \sum_{i=1}^{N} P_{Kig}(\hat{P}'_{Kig}\hat{\theta} - P'_{Kig}\theta) \right\| = O_{\mathbb{P}}(\zeta_K^2 \sqrt{K/G} + \zeta_K \sqrt{K/N})$$
$$\left\| \frac{1}{N} \sum_{i=1}^{N} P_{Kig} r_{ig} \right\| = O_{\mathbb{P}}(\zeta_K K^{-\alpha}).$$

Thus, $\|\hat{\psi}_{Ng} - \psi_{Ng}\| = O_{\mathbb{P}}(\zeta_K(\zeta_K\sqrt{K/G} + \sqrt{K/N} + K^{-\alpha}))$ and the conclusion follows from $\|\psi_{Ng}\| + \|\hat{\psi}_{Ng}\| = O_{\mathbb{P}}(\zeta_K)$. □

Lemma 8 (Lemma 2.2 in Schmitt (1992))**.** *Let A and B two symmetric matrices satisfying $A \succ \mu_a^2 \mathrm{I}$, $B \succ \mu_b^2 \mathrm{I}$, where $A \succ B$ denotes $A - B$ is positive definite. Then,*

$$\|A^{1/2} - B^{1/2}\| \leq \{\mu_a + \mu_b\}^{-1} \|A - B\|.$$

A.5 Proof of Theorem 4

By taking conditional expectation give $F_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g)$,

$$\mathbb{E}[Y_{iJ_{ig}g}|F_{W|J,\mathbf{B}}] = \mathbb{E}[X_{J_{ig}g}|F_{W|J,\mathbf{B}}]'\beta + \mathbb{E}[W_{ig}|F_{W|J,\mathbf{B}}]'\gamma + \int \psi(w) dF_{W|J,\mathbf{B}}(w|J_{ig}, \mathbf{B}_g).$$

Then, subtracting it from the original equation,

$$Y_{iJ_{ig}g} - \mathbb{E}[Y_{iJ_{ig}g}|F_{W|J,\mathbf{B}}] = \{V_{ig} - \mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}}]\}' \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \varepsilon_{ig}$$

where $V_{ig} = (X'_{J_{ig}g}, W'_{ig})'$. Then, letting $\tilde{Z}_{ig} = (Z'_{J_{ig}g}, W'_{ig})'$,

$$\mathbb{E}[\tilde{Z}_{ig}\{Y_{iJ_{ig}g} - \mathbb{E}[Y_{iJ_{ig}g}|F_{W|J,\mathbf{B}}]\}] = \mathbb{E}[\tilde{Z}_{ig}\{V_{ig} - \mathbb{E}[V_{ig}|F_{W|J,\mathbf{B}}]\}'] \begin{pmatrix} \beta \\ \gamma \end{pmatrix}$$

and full column rank of the matrix implies identifiability of β.

A.6 Proof of Theorem 5

By the independence assumption

$$\begin{aligned} f_{W,\varepsilon|J,\mathbf{B}}(w,e|s,\mathbf{b}) &= \int f_{W,\varepsilon|J,\mathbf{B},\Theta}(w,e|s,\mathbf{b},\theta) f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\theta \\ &= \int f_{W,\varepsilon|\Theta}(w,e|\theta) f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\theta \end{aligned}$$

and similarly,

$$f_{W|J,\mathbf{B}}(w|s,\mathbf{b}) = \int f_{W|\Theta}(w|\theta) f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}) d\theta.$$

By injectivity of $\mathbf{\Psi}$, we have

$$\mathbf{\Psi}^{\dagger}\big(f_{W|J,\mathbf{B}}(\cdot|s,\mathbf{b})\big)(\theta) = f_{\Theta|J,\mathbf{B}}(\theta|s,\mathbf{b}).$$

Then, conditioning on $F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)$, the function $f_{W,\varepsilon|J,\mathbf{B}}$ is non-stochastic. Therefore,

$$\begin{aligned} &\mathbb{E}[Y_{iJ_{ig}g}|X_{J_{ig}g}=x, F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)] \\ &= \mathbb{E}[m(x,W_{ig},\varepsilon_{iJ_{ig}})|X_{J_{ig}g}=x, F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)] \\ &= \mathbb{E}[\mathbb{E}[m(x,W_{ig},\varepsilon_{iJ_{ig}})|X_{J_{ig}g}=x, J_{ig},\mathbf{B}_g]|X_{J_{ig}g}=x, F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)] \\ &= \mathbb{E}\Big[\int m(x,w,e)\mathbb{E}[f_{W,\varepsilon|J,\mathbf{B}}(w,e|J_{ig},\mathbf{B}_g)|X_{J_{ig}g}=x,\mathbf{B}_g]\big|X_{J_{ig}g}=x, F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)\Big] \\ &= \int m(x,w,e)\mathbb{E}[f_{W,\varepsilon|J,\mathbf{B}}(w,e|J_{ig},\mathbf{B}_g)|X_{J_{ig}g}=x, F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]d(w,e) \\ &= \int m(x,w,e)\mathbb{E}[f_{W,\varepsilon|J,\mathbf{B}}(w,e|J_{ig},\mathbf{B}_g)|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g)]d(w,e) \end{aligned}$$

where the third equality use the display below, the second-to-last equality uses the Fubini theorem to interchange the order of integration, and the last equality uses $f_{W\varepsilon|J,\mathbf{B}}$ is non-stochastic conditional on $F_{W|J,\mathbf{B}}$. For the third equality in the above display,

$$\begin{aligned} \mathbb{E}[m(x,W_{ig},\varepsilon_{iJ_{ig}g})|J_{ig},\mathbf{B}_g] &= \sum_{s=1}^{S}\mathbb{E}[m(x,W_{ig},\varepsilon_{isg})|J_{ig}=s,\mathbf{B}_g]\Pr(J_{ig}=s|\mathbf{B}_g) \\ &= \sum_{s=1}^{S}\int m(x,w,e) f_{W,\varepsilon|J,\mathbf{B}}(w,e|s,\mathbf{B}_g)d(w,e)\Pr(J_{ig}=s|\mathbf{B}_g) \\ &= \int m(x,w,e)\mathbb{E}[f_{W,\varepsilon|J,\mathbf{B}}(w,e|J_{ig},\mathbf{B}_g)|\mathbf{B}_g]d(w,e). \end{aligned}$$

Finally, letting ν be the measure corresponding to the distribution of $F_{W|J,\mathbf{B}}$, which is identifiable from the data,

$$\begin{aligned}
&\int \mathbb{E}[Y_{iJ_{ig}g}|X_{J_{ig}g} = x, F_{W|J,\mathbf{B}}(\cdot|J_{ig}, \mathbf{B}_g) = F]d\nu(F) \\
&= \int m(x,w,e)\mathbb{E}[f_{W,\varepsilon|J,\mathbf{B}}(w,e|J_{ig},\mathbf{B}_g)|F_{W|J,\mathbf{B}}(\cdot|J_{ig},\mathbf{B}_g) = F]d(w,e)d\nu(F) \\
&= \int m(x,w,e)\mathbb{E}[f_{W,\varepsilon|J,\mathbf{B}}(w,e|J_{ig},\mathbf{B}_g)]d(w,e) \\
&= \int m(x,w,e)f_{W,\varepsilon}(w,e)d(w,e).
\end{aligned}$$

B Details of DGPs for Monte Carlo Studies

The details of the DGPs used in the simulation studies are following. I use $G = 1,000$ and for each city, $N = 300$ and $S = 3$ with the constraint that each group can have at most 115 people. The final sample from the NLS72 dataset contains 917 schools, which I treated as independent draws. This assumption is valid if schools are geographically isolated from each other.

The econometric model generating the data is

$$\begin{aligned}
Y_{is} &= X_s\beta + W_i'\gamma + \omega_i + \chi_s + \epsilon_{is} \\
J_i &= \underset{s\in\{1,2,3\}}{\arg\max}\{\Theta_i'A_s + \eta_{is}\}
\end{aligned}$$

where $\beta = -0.15$ and $\gamma = (-0.0003, 0.06)'$ are taken from an estimate in the NLS72. Also, W_i has a bivariate normal distribution with mean $(12.37, 10.92)$ and covariance matrix

$$\begin{bmatrix} 4.25 & 0.49 \\ 0.49 & 0.42 \end{bmatrix}.$$

The means and covariance matrix are based on the NLS72 data. The preference coefficient Θ_i is generated by

$$\Theta_i = W_i + \omega_i\mathbf{1} + \nu_i$$

where $\mathbf{1}$ is 2×1 vector whose elements are unity and ν_i is a mean-zero bivariate normal with covariance matrix $\begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix}$. For χ_s and ϵ_{is}, they have a mean-zero normal distribution with standard deviation 0.446, which imitates the standard error of a regression from the data. The

idiosyncratic term η_{is} is drawn from mean-zero normal distribution with standard deviation 2. For X_s, I use the empirical distribution of a school-level variable observed in the NLS72 dataset; A_s consists of X_s and a binary variable that takes value 1 with probability 0.29 and has correlation of 0.25 with X_s.

For ω_i, each DGP generates this variable in different ways. For DGP1, ω_i has a joint normal distribution with W_i, its mean is 2.16, variance is 0.51, and the covariance with W_i is 0.051, 0.024. The joint normality leads to the linear specification of the control function. For DGP2, I model

$$\omega_i = 0.1 * u_i + 2 * \mathbb{1}\{W_{1i} + W_{2i} - \mathbb{E}[W_{1i} + W_{2i}] > v_i\}$$

where u_i has the same distribution as ω_i in DGP1 (jointly normal with W_i) and $v_i =_d \text{Normal}(0, 0.2)$. For the third DGP,

$$\omega_i = 0.5 * W_{1i}^2 + 0.5 * W_{2i}^3$$

where W_{1i} and W_{2i} represent the first and second component of W_i.

In terms of implementation, I generate the random variables and compute utilities of choosing different groups for each individual. Then, I use linear programming to find an allocation that maximizes the sum of utilities under the group size constraint. Altonji and Mansfield use this step in their simulation and I follow their procedure.