Peer Effects with Sample Selection[*]

Xin Gu[†] Haizheng Li[‡] Zhongjian Lin [§] Xun Tang[¶]

May 3, 2024

Abstract

We study peer effects in social interaction models where individual participation in groups is endogenous. By exploiting instruments in participation choices, we estimate the peer and contextual effects using both individual and group-level correction terms. We apply our method to study peer effects in an online training program in China, and document significant peer effects and selection bias in the duration of lecture attendance among the trainees. Our estimates suggest that ignoring sample selection would over-estimate the peer effects.

Keywords: Peer Effects, Sample Selection, Social Interactions, Reflection Problem, Online Training.

JEL Codes: C24, C31, C36, M53.

[*] We are grateful to Daniel Dench, Peng Shao, Nicholas Ziebarth, and the audience at Auburn University for useful comments and suggestions. We thank the YouChange China Social Entrepreneur Foundation for providing the data. All errors are our own.

[†] Xi'an Jiaotong-Liverpool University

[‡] Georgia Institute of Technology

[§] University of Georgia

[¶] Rice University

1 Introduction

Sample selection exists in many social-economic contexts (Heckman, 1974, 1978, 1979). A typical example of such context is the wage structure of female workers, where wages are only observed for those participating in the labor force (Gronau, 1974, Heckman, 1974). A classical approach for dealing with sample selection is the two-step correction method in (Heckman, 1979), which requires parametric assumptions on the correlation between unobserved errors determining sample selection and observed outcomes. This approach allows for dependence between selection and outcomes within each independent observation, but can not be applied when there are spillover (e.g., peer or contextual) effects in the outcome and/or selection across individuals.

For instance, in the example above, the workers' outcomes (wages) may well be determined simultaneously within small groups (e.g., those in the labor force from the same geographic area) as in a social interactions model with peer, contextual and correlated effects. Even if a worker's decision to participate in the labor force depends solely on her own idiosyncratic factors, a proper correction for such selection bias would require dealing with the endogenous selection of *the other workers in the same group* as well. This is simply because in the presence of social interactions, the reduced-form of each worker's wage also depends on the idiosyncratic factors of *other workers* in the same group. As a result, the selection bias in these other group peers also need to be addressed properly in the reduced form of individual outcomes.

In this paper, we investigate the sample selection issue in a social interactions model where individual outcomes are influenced by peer effects. Social interactions models with peer effects have proliferated in empirical research in recent decades (Manski, 1993, 2000, Brock and Durlauf, 2001a, Moffitt, 2001, Lee, 2007, Bramoullé, Djebbari, and Fortin, 2009). Evidence of peer effects has been found in many fields, including economics of education (Hoxby, 2000, Sacerdote, 2001, Calvó-Armengol, Patacchini, and Zenou, 2009), financial economics (Hong, Kubik, and Stein, 2004), health economics (Trogdon, Nonnemaker, and Pais, 2008), labor economics (Topa, 2001, Dahl, Løken, and Mogstad, 2014), and urban economics (Glaeser, Sacerdote, and Scheinkman, 1996).

Most existing empirical papers that study peer effects use experiments or quasi-experiments to randomly assign members into groups. In such cases, group formation can be taken as exogenously given. In contrast, we explain group formation in a social interactions model by allowing *potential* members of a group to endogenously join the *actual* group (e.g., participate in the labor force from the same area). We take the definition of *potential* groups as given and fixed, so that self-selection arises only due to individual decisions to participate in the *actual* groups. The outcome for each actual group member is then affected by, and determined simultaneously with, the outcomes of group peers. Again, consider the wage example above. In that setting, a potential group consists of individuals from the same area who could *potentially* join the labor force. The actual group then consists of participants in the labor force from that area who were recorded in the sample with wage information.

Several earlier papers studied peer effect models with other forms of endogeneity in group memberships. Carrell, Sacerdote, and West (2013) showed that group composition can be highly endogenous in practice. Goldsmith-Pinkham and Imbens (2013), Hsieh and Lee (2016), and Auerbach (2022) studied network models with dyadic links formed from unobserved individual heterogeneity. Boucher (2016) considered a model of conformism where agents simultaneously choose a continuous outcome and which peers to form links with. Badev (2021) developed a model where individuals simultaneously make binary decisions and choose the links to form. Johnsson and Moon (2021) and Jochmans (2023) studied the peer effects when the social network is endogenous to the outcomes due to unobserved individual characteristics. Sheng and Sun (2021) modeled group formation using a notion of stability in a matching model.

In contrast with these papers, we deal with a qualitatively different form of endogeneity in group formation. In our case, individuals in exogenously-defined *potential* groups choose whether to actively join the *actual* group. Specifically, we consider an application of online training program for elementary and middle school teachers in rural China. In this context, we define a potential group as the set of all teachers enrolled in the training program from the same county. We then construct an actual group as the subset of the potential group, who attended a particular lecture. In other words, self-selection into an actual group simply means the teacher chose to attend that lecture. The outcome of interest is the duration for which a teacher stayed in a lecture. These individual outcomes are continuous, simultaneously determined, and only reported for the actual group members. We combine the identification of social interactions models with the solution to sample selection issues.

Our identification results are due to earlier ideas from Manski (1993) and Brock and Durlauf (2001b). To put this in context, consider a linear-in-means social interactions model with a contextual effect from the characteristics of group members, and an endogenous peer effect capturing a structural simultaneity between all individual outcomes within a group.[1] Manski (1993) pointed out that, without further restrictions, the peer effects can not be separated from the contextual effects from the reduced form coefficients in this model. He proposed a solution to the problem using an exclusion restriction, i.e. there are covariates with non-trivial direct effects but no contextual effects (Proposition 2). Similar exclusion restrictions were used in Moffitt (2001) for identifying peer effects.[2] The identification strategy we use in this paper is based on insights from

[1] In the terminology of social interactions, contextual effects refer to how the characteristics of group peers directly impact an individual's outcome in a structural form. In comparison, peer effects reflect how the individual outcomes of *all* group peers are jointly determined in a simultaneous system.

[2] There are other alternatives for solving the identification problem: second-moment restrictions on the error terms (Lee, 2007, Graham, 2008, Sacerdote, 2001), variation in the group sizes (Bramoullé, Djebbari, and Fortin, 2009, De Giorgi, Pellizzari, and Redaelli, 2010, Lin, 2010), and control functions for endogenous covariates (Lin and Tang, 2022), etc.

Brock and Durlauf (2001b) (Section 3.6). Hoshino (2019) studies social interactions with incomplete information and missing observations due to sample selection. The method, however, is different from this paper. He uses a different identification strategy, i.e., network variation (Bramoullé, Djebbari, and Fortin, 2009) and proposes a two-step series nonlinear least squares estimator.

We propose a multi-step estimator for peer and contextual effects, building on this constructive argument for identification. First, a Probit regression of the binary response model (a.k.a. the selection equation) provides consistent estimates for the selection correction term. Second, by including the estimates of the individual and group correction terms as generated regressors in a linear regression, we estimate the reduced-form coefficients in a social interactions model, which are then used for backing out the peer and contextual effects.

We generalize the core idea in Section 4 to allow group-level unobserved heterogeneity (GUH) in outcome as well as group selection. Both are empirically relevant extensions. In the case with GUH in outcomes, we use approaches analogous to fixed-effect or random-effect methods in panel data. The case with GUH in group selection is more complicated, and requires a correlated random-effect method which entails maximum simulated likelihood in the first step.

We then study the peer effects in an online training program for elementary and middle school teachers in China. In this setting, a *potential* group consists of teachers enrolled in the program from the same county. These teachers decided to participate or skip each specific lecture, based on self-motivation and other factors. This results

Brock and Durlauf (2001b) (Section 3.6) showed such exclusion restrictions naturally arise when individuals' endogenous self-selection into groups are accounted for in social interaction models. They considered a setting where each individual is associated with a *reservation* group, and rationalized individual participation in observed groups through binary choices. They exploited the exogenous variation in individual instruments in binary responses to identify the peer effects. Brock and Durlauf (2001b) noted that the individual correction terms for self-selection essentially could essentially function as *generated* regressors which satisfy the exclusion restriction in Manski (1993). Consequently, peer and contextual effects can be recovered from coefficients in the reduced-form equation. See Equation (69) and (71) in Brock and Durlauf (2001b) for details.

in a non-random sample of actual groups (lecture attendants from the same county) based on endogenous self-selection. The outcome of interest is the duration of lecture attendance by each individual. This depends on unobserved noises related to self-motivation, which may well be correlated with those determining lecture participation in the first place. For example, the duration of lecture attendance is specified as a model with peer and contextual effects, while its error may be related to the decision on participation (self-selection). Thus, this endogenous sample selection gives rise to new challenges for estimating peer effects in the duration of lecture attendance. Using our multi-step estimator, we find significant peer effects among trainees attending the same lecture. Also, ignoring the sample selection issue in this context would result in an erroneous conclusion about the magnitude and significance of the peer effects. Ignoring sample selection would over-estimate the peer effects. Teachers may decide to stay longer because they are more motivated, not only because of positive peer effects. So, if we drop the first factor (motivation) in estimation by ignoring sample selection, we may end up over-stating the significance of the latter factor (peer effects).

It is worth emphasizing that the method in this paper provides a feasible way to study the effect of policy interventions under social interactions when there is imperfect compliance with group assignment (i.e., researchers have less control over the group composition). In randomized control trials (RCTs), a researcher may offer exogenous incentives for group participation. For instance, one may introduce random factors that encourage group participation but do not directly affect individual outcomes (e.g., by offering credits or subsidies for lecture attendants). Such random factors could then serve as instruments in the selection equation; their variation across the groups helps researchers to identify peer and contextual effects.

The paper unfolds as follows. We introduce the peer effects model with sample selection and discuss its identification in the next section. We propose a multi-step estimator in Section 3. Section 4 then shows how to extend the method where there are unobserved group fixed effects. We show the finite-sample performance of the estimator via Monte Carlo simulations in Section 5. Finally, we apply our method in the empirical application of peer effects in the online training program for teachers in Section 6. Section 7 concludes.

2 The Model

We consider a data-generating process (DGP) which generates a large number of independent groups, indexed by $g = 1, 2, ..., G$. Each group has a set of *potential* members, denoted by $\mathscr{N}_g$. We suppress the group index g in the notation in this section. Our model specification is similar to that considered in Section 3.6 of Brock and Durlauf (2001b), but deviates by allowing the peer and contextual effects to operate through group averages, as opposed to some (conditional) population expectation.[3]

The model extends conventional linear-in-means social interactions models by allowing individuals to join a group as *actual* members through endogenous self-selection (e.g., teachers enrolled in the training program from the same county decide whether to attend a specific lecture). Specifically, for $i \in \mathscr{N}$, the decision to join a group is determined as:

$$D_i = 1\{Z_i'\delta + V_i \geq 0\}. \tag{1}$$

Henceforth we refer to Z_i as individual *instruments*.

For each group, let n denote the number of actual members, i.e., $\{i \in \mathscr{N} : D_i = 1\}$.[4] The vector of outcomes (e.g., the length of time for which a teacher stayed in the lecture)

[3] In the latter case, a consistent estimation of the group-level correction term, i.e., $E[\lambda(\gamma' R_i)|i \in n(i)]$ in Brock and Durlauf (2001b), would not be practical if the group sizes are small and the individual labels can not be matched across the groups.

[4] Without loss of generality, we index the members of an actual group as $i = 1, 2, ..., n$.

of actual group members are determined simultaneously as:

$$Y = \alpha\overline{Y} + \beta_0 + X'\beta + \overline{X}'\gamma + U, \tag{2}$$

where X' is an $n \times K$ matrix of individual characteristics (which does not include a constant intercept), $\overline{X}'$ is an $n \times K$ matrix of n identical rows with each being a $1 \times K$ vector of average characteristics within the group, $\overline{Y}$ is the average outcome of individuals within the group, and $U \equiv (U_i)_{i \leq n}$ is an $n \times 1$ vector of individual structural errors. We use an overall group average, as opposed to a "leave-one-out" average of *other* peers. This is the empirical analog of the original linear-in-mean specification in Manski (1993). This specification is used in a variety of empirical contexts (see, for example, Trogdon, Nonnemaker, and Pais, 2008, Mora and Gil, 2013).

The individual instruments Z_i contain elements that are not in X_i. While the data report (D_i, Z_i) for all *potential* group members $i \in \mathscr{N}$, it only reports individual outcomes Y_i and demographics X_i for n *actual* group members.

By solving for the reduced form of $\overline{Y}$ and substituting it back into the structural form in (2), we have

$$Y = \tilde{\beta}_0 + X'\beta + \overline{X}'\tilde{\gamma} + \widetilde{U}, \tag{3}$$

where $\widetilde{U} \equiv U + \frac{\alpha}{1-\alpha}\overline{U}$ with $\overline{U} \equiv \frac{1}{n}\sum_{i \leq n} U_i$, $\tilde{\beta}_0 \equiv \frac{\beta_0}{1-\alpha}$, and $\tilde{\gamma} \equiv \frac{\alpha\beta+\gamma}{1-\alpha}$. Let $\mathscr{S}$ be shorthand for the selection event that "$D_i = 1$ for all $i = 1, 2, ..., n$ and $D_j = 0$ for all other $j \in \mathscr{N}$". Define $\varepsilon \equiv \widetilde{U} - E(\widetilde{U}|X, Z, \mathscr{S})$, where $Z \equiv (Z_i)_{i \in \mathscr{N}}$ denotes the vector of all instruments associated with all potential members. Then write (3) as

$$Y = \tilde{\beta}_0 + X'\beta + \overline{X}'\tilde{\gamma} + E(\widetilde{U}|X, Z, \mathscr{S}) + \varepsilon, \tag{4}$$

where (X, Z) are exogenous in the sense that $E(\varepsilon|X, Z, \mathscr{S}) = 0$. Thus the conditional mean of $\widetilde{U}$ in (4) serves as a correction for the sample selection bias, which is due to the correlation between U and $V \equiv (V_i)_{i \in \mathscr{N}}$ (e.g., the unobserved factors affecting the duration of a teacher's attendance is correlated with those in his own and others' decisions on whether to attend the lecture). We maintain the following assumptions, which allow us to derive the correction term.

Assumption 1. *(i) $E[U_i|V,Z,(X_i)_{i\in\mathscr{N}}] = E(U_i|V_i)$ for each $i \in \mathscr{N}$. (ii) V is independent from $(X_i,Z_i)_{i\in\mathscr{N}}$, and V_i's are independent across $i \in \mathscr{N}$. (iii) For each i, the vector (U_i,V_i) follows a bivariate normal distribution with $\sigma_{uv} \neq 0$:*

$$\begin{pmatrix} U_i \\ V_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma_{uv} \\ \sigma_{uv} & 1 \end{pmatrix} \right).$$

Remark 1. *Assumption 1 allows the n-vector of structural errors U to be correlated among individual members, possibly through group fixed effects. It only restricts the correlation between U_i and V_i for each i, as well as the independence between the selection errors V_i across $i \in \mathscr{N}$.*[5] *In Section 4.2 we generalize the setting by allowing individual unobserved heterogeneity to be correlated through group fixed effects. Note that the issue of sample selection arises because the structural errors in the outcomes and participation decisions are correlated, i.e., $\sigma_{uv} \neq 0$. If $\sigma_{uv} = 0$ in Assumption 1, there would be no sample selection bias at all.*

Under Assumption 1, we have

$$E(U_i|X,Z,\mathscr{S}) = E(U_i|V_i \geq -Z_i'\delta) = \sigma_{uv}\lambda(Z_i'\delta), \tag{5}$$

where $\lambda(Z_i'\delta) = \frac{\phi(-Z_i'\delta)}{1-\Phi(-Z_i'\delta)} \equiv \lambda_i$ is the inverse Mills ratio (as in Heckman, 1979). Since its introduction by Heckman (1979), this method for correcting sample selection bias has been applied widely in theory and practice. After correcting for the selection bias, we write the reduced form for each i as

$$Y_i = \tilde{\beta}_0 + X_i'\beta + \overline{X}'\tilde{\gamma} + \sigma_{uv}\lambda_i + \tilde{\sigma}_{uv}\bar{\lambda} + \varepsilon_i \tag{6}$$

[5] For example, suppose (U,V) is multivariate normal and independent from (X,Z). Assumption 1 does not restrict the off-diagonal entries in the upper-left quadrant of its covariance matrix. On the other hand, it requires all three of the remaining quadrants to be diagonal.

where $\tilde{\sigma}_{uv} \equiv \frac{\alpha}{1-\alpha}\sigma_{uv}$, and $\bar{\lambda}$ is the average of λ_i over all $i \leq n$ in a group.

Let $W_i \equiv (1, X_i', \overline{X}', \lambda_i, \bar{\lambda})$ denote a row-vector in R^{2K+3}. As long as $E(W_i'W_i)$ is non-singular, we can consistently estimate $(\tilde{\beta}_0, \beta, \tilde{\gamma}, \sigma_{uv}, \tilde{\sigma}_{uv})$ by regressing Y_i on W_i. Then we can recover the structural parameters in (2) as:

$$\alpha = \frac{\tilde{\sigma}_{uv}}{\tilde{\sigma}_{uv} + \sigma_{uv}}; \ \beta_0 = (1-\alpha)\tilde{\beta}_0; \ \gamma = (1-\alpha)\tilde{\gamma} - \alpha\beta. \tag{7}$$

As noted above, if $\sigma_{uv} = 0$, there would be no sample selection bias, and the model of outcomes in (2) would reduce to a standard linear-in-means model. In that case, we have $2K+1$ parameters, $\{\tilde{\beta}_0, \beta, \tilde{\gamma}\}$ in the reduced form and $2K+2$ parameters $\{\alpha, \beta_0, \beta, \gamma\}$ in the structural form. The failure of order condition – there are fewer reduced-form coefficients than structural ones – leads to the reflection problem (Manski, 1993). With $\sigma_{uv} \neq 0$, we introduce one additional parameter σ_{uv} in the structural form, but generate two parameters $\{\sigma_{uv}, \tilde{\sigma}_{uv}\}$ in the reduced form. Thus, the number of structural and reduced-form parameters are equal. Furthermore, the rank condition required for recovering the structural parameters from the reduced form ones also holds. Without sample selection, we would not be able to use individual instruments from (1) as a source of exogenous variation to help us resolve the reflection problem.

There is another intuitive interpretation of our method. We have introduced individual-level correction terms for the selection bias in the structural form of this model. These correction terms then conveniently function as *generated* regressors which satisfy the exclusion restrictions in Manski (1993). Therefore, we are able to solve the "reflection problem" without imposing further assumptions.

3 Estimation

We define a multi-step estimator based on the constructive identification strategy above. For simplicity, we present the estimator when X is a strict sub-vector of Z; generalization to cases where Z contains distinct elements from X is straightforward.

Let the sample contain G independent groups. Each group g is formed out of a finite superset of *potential* members, which is denoted by $\mathscr{N}_g$. Each potential member $i \in \mathscr{N}_g$ chooses to join the group g or not, $D_{g,i} \in \{0,1\}$, by following the rule in Equation (1). We refer to those who choose $D_{g,i} = 1$ and self-select into the group as the *actual* group members. Let $n_g = \sum_{i \in \mathscr{N}_g} D_{g,i}$ denote the actual size of group g. For each group $g \leq G$, the sample reports $(D_{g,i}, Z_{g,i})$ for all potential members $i \in \mathscr{N}_g$, but only reports $Y_{g,i}$ for actual group members who self-select to join the group $(D_{g,i} = 1)$. Similar to Section 2, let $\mathscr{S}_g$ denote the sample selection event in potential group g.

The identification strategy in Section 2 applies to groups with at least two actual members. Formally, this means the sample correction term in Equation (4) conditions on $n_g \geq 2$. The identification strategy in Section 2 applies because under Assumption 1 the individual correction term takes the form in Equation (5). That is, $E(U_{g,i} | X_g, Z_g, \mathscr{S}_g, n_g \geq 2) = E(U_{g,i} | V_{g,i} \geq -Z'_{g,i}\delta)$.

The first step of our estimator is to construct individual correction terms $\lambda_{g,i}$'s for $i \leq n_g$ by running a Probit regression of $D_{g,i}$ on $Z_{g,i}$ in Equation (1) using *all* potential group members $i \in \mathscr{N}_g$. Let $\hat{\delta}$ denote the Probit estimator for δ from this step. For each actual member $i \leq n_g$ in group g, calculate

$$\hat{\lambda}_{g,i} \equiv \phi(Z'_{g,i}\hat{\delta}) / \Phi(Z'_{g,i}\hat{\delta}), \text{ and } \hat{\bar{\lambda}}_g \equiv \frac{1}{n_g} \sum\nolimits_{i=1}^{n_g} \hat{\lambda}_{g,i}.$$

The second step is an OLS regression of $Y_{g,i}$ on $X_{g,i}, \overline{X}_g, \hat{\lambda}_{g,i}$ and $\hat{\bar{\lambda}}_g$ using the actual group members. Let $\theta \equiv \left(\tilde{\beta}_0, \beta', \tilde{\gamma}', \sigma_{uv}, \tilde{\sigma}_{uv}\right)'$ be a column-vector that collects all reduced-form parameters to be estimated in this step. For each group g and actual group member $i \leq n_g$, define a row-vector of generated regressors:

$$W_{g,i}(\hat{\delta}) \equiv \left(1, X'_{g,i}, \overline{X}'_g, \hat{\lambda}_{g,i}, \hat{\bar{\lambda}}_g\right).$$

Denote the total number of actual group members in the sample by $N = \sum_{g \leq G} n_g$. Let

$W(\hat{\delta})$ be an N-by-dim(θ) matrix that stacks the row-vector of generated regressors $W_{g,i}(\hat{\delta})$ from all groups and actual members, and Y be an N-by-1 vector that stacks the column-vectors Y_g from all groups in the sample. Our estimator for θ in this step is constructed by regressing Y on $W(\hat{\delta})$:

$$\hat{\theta} \equiv \left[\sum\nolimits_{g,i} W_{g,i}(\hat{\delta})' W_{g,i}(\hat{\delta})\right]^{-1} \left[\sum\nolimits_{g,i} W_{g,i}(\hat{\delta})' Y_{g,i}\right] = [W(\hat{\delta})' W(\hat{\delta})]^{-1} W(\hat{\delta}) Y,$$

where $\sum_{g,i}$ is shorthand for the double summation $\sum_{g \leq G} \sum_{i \leq n_g}$. By definition, $\hat{\theta}$ provides estimators for the reduced-form parameters $(\hat{\tilde{\beta}}_0, \hat{\beta}, \hat{\tilde{\gamma}}, \hat{\sigma}_{uv}, \hat{\tilde{\sigma}}_{uv})$.

The last step is to calculate the structural parameters from $\hat{\theta}$ using Equation (7). Denote these estimators by $(\hat{\alpha}, \hat{\beta}_0, \hat{\gamma})$.

We sketch a proof for the asymptotic property of the two-step m-estimator $\hat{\theta}$. Let $A \equiv \lim_{G \to \infty} \frac{1}{G} \sum_{g \leq G} E\left(W_g' W_g\right)$, where W_g is shorthand for $W_g(\delta)$, which stacks $W_{g,i}(\delta)$ over all $i \leq n_g$, and is evaluated at the true parameter δ in the selection equation (1). First, under standard regularity conditions, e.g., including finite, non-singular A and those in Lemma 4.3 of Newey and McFadden (1994), $\frac{1}{G} W(\hat{\delta})' W(\hat{\delta})$ and $\frac{1}{G} W(\hat{\delta})' Y$ converge in probability to A and $A\theta$ respectively as $G \to \infty$. With A invertible, this implies consistency of the estimator: $\hat{\theta} \xrightarrow{p} \theta$.

Next, under standard regularity conditions, the first-order condition in the second-step regression implies:

$$\sqrt{G}\left(\hat{\theta} - \theta\right) = A^{-1} \left\{-G^{-1/2} \sum\nolimits_g s_g(\theta; \hat{\delta})\right\} + o_p(1),$$

where $s_g(\theta; \hat{\delta}) \equiv W_g(\hat{\delta})' \left[Y_g - W_g(\hat{\delta})\theta\right]$, with $W_g(\hat{\delta})$ being n_g-by-dim(θ) and stacking $W_{g,i}(\hat{\delta})$ across i in each group g. A second-order mean-value expansion of $s_g(\theta; \hat{\delta})$ around δ implies:

$$G^{-1/2} \sum\nolimits_g s_g(\theta; \hat{\delta}) = G^{-1/2} \sum\nolimits_g s_g(\theta; \delta) + F_{g,0} \sqrt{G}(\hat{\delta} - \delta) + o_p(1),$$

where $F_{g,0} \equiv E\left[\nabla_\delta s_g(\theta; \delta)\right] \in \mathbb{R}^{dim(\theta) \times dim(\delta)}$. Let $r_g(\delta)$ denote the influence function in the asymptotic linear representation of the first-step estimator $\hat{\delta}$. That is, $\sqrt{G}(\hat{\delta}-\delta)=G^{-1/2}\sum_g r_g(\delta)+o_p(1)$. The limiting distribution of $\hat{\theta}$ is then:

$$\sqrt{G}\left(\hat{\theta}-\theta\right)\xrightarrow{d}\mathscr{N}(0,A^{-1}BA^{-1}),$$

where $B\equiv\lim_{G\to\infty}\frac{1}{G}\sum_{g\leq G}E[m_g(\theta;\delta)m_g(\theta;\delta)']$, with $m_g(\theta;\delta)\equiv s_g(\theta;\delta)+F_{g,0}r_g(\delta)$.

The components in asymptotic variance A,B can both be consistently estimated by plugging parameter estimates into their respective sample analogs. In our empirical application, we use bootstrap resampling methods to calculate the standard errors.

In the last step, the remaining structural parameters, i.e., the peer effect α, the contextual effects γ and the intercept β_0 are estimated by plugging $\hat{\theta}$ in the formulas in (7). Asymptotic variance of these structural parameters can be obtained by a direct application of the Delta Method.

4 Extensions: Group Fixed Effects

4.1 Group Fixed Effects in Outcomes

We extend our method to allow for unobserved group fixed effects in the outcomes:

$$Y=\alpha\overline{Y}+\beta_0+X'\beta+\overline{X}'\gamma+\eta+U, \tag{8}$$

where η is an unobserved group-level fixed effect. As before, individual selection into the sample is based on Equation (1).

Using the same restrictions on (U,V) as Assumption 1 in Section 2, we get

$$Y=\tilde{\beta}_0+X\beta+\overline{X}\tilde{\gamma}+\sigma_{uv}\lambda+\tilde{\sigma}_{uv}\overline{\lambda}+\tilde{\eta}+\varepsilon, \tag{9}$$

where $\tilde{\eta}\equiv\eta/(1-\alpha)$ and $(\tilde{\beta}_0,\tilde{\gamma},\varepsilon)$ are defined as in (4) so that $E(\varepsilon|X,Z,\mathscr{S})=0$, which implies (X,Z) are exogenous in the sample. Without further restrictions, $\tilde{\eta}$ is generally correlated with (X,Z,U,V) and therefore ε.

There are several ways to estimate the parameters in (9), depending on the assumption on how η is correlated with the other variables. First, in the simplest case, suppose η is independent from (U,V,Z,X) and therefore ε. Write $\tilde{\eta}+\varepsilon = E(\tilde{\eta})+\tilde{\varepsilon}$ with $\tilde{\varepsilon} \equiv \varepsilon + \tilde{\eta} - E(\tilde{\eta})$, so that $E(\tilde{\varepsilon}|X,Z,\mathscr{S}) = 0$. By regressing y on $(1,X,\overline{X},\lambda,\overline{\lambda})$ and using a necessary location normalization $E(\tilde{\eta}) = 0$, we can consistently estimate $(\tilde{\beta}_0,\beta',\tilde{\gamma}',\sigma_{uv},\tilde{\sigma}_{uv})$. This then identifies (α,β_0,γ).

In a setting where η is correlated with (U,V,X,Z), one can estimate the model using instruments, i.e., additional observed variables that are uncorrelated with η but are correlated with (X,Z), thus satisfying instrument exogeneity and relevance. Using these instruments, one can consistently estimate $(\tilde{\beta}_0,\beta',\tilde{\gamma}',\sigma_{uv},\tilde{\sigma}_{uv})$ in (9) through two-stage least squares.

Yet another option for estimating the model with correlated fixed effects is to parameterize the dependence between η and (U,V), and construct a correction term. In the next section, we elaborate on this solution in more general settings.

4.2 Group Fixed Effects in Sample Selection

We now generalize the model in Section 4.1 by accommodating a second unobserved group fixed effect in the sample selection equation. Suppose individuals self-select into a group in the sample as follows:

$$D_i = 1\{Z_i\delta + \zeta + V_i \geq 0\}, \tag{10}$$

for each potential group member $i \in \mathscr{N}$, where ζ is an unobserved group fixed effect. The structural form of outcomes within each group is the same as (8), which includes a group fixed effect η.

For convenience, define $V_i^* \equiv \zeta + V_i$ and $U_i^* = \eta + U_i$; let $V^* \equiv (V_i^*)_{i\in\mathscr{N}}$ and $Z \equiv (Z_i)_{i\in\mathscr{N}}$; let $U^* \equiv (U_i^*)_{i\leq n}$ and $X \equiv (X_i)_{i\leq n}$, where n denotes the number of actual group members i with $D_i = 1$. We maintain the following assumptions.

Assumption 1'. *(i) V is independent from (ζ, Z), and V_i are independently distributed as standard normal across potential group members. (ii) $(U_i^*)_{i\in\mathscr{N}}$ and V^* are independent from $(X_i)_{i\in\mathscr{N}}$ and Z. (iii) $E(U_i^*|V,\zeta) = E(U_i^*|V_i,\zeta) = \pi_1\zeta + \pi_2 V_i$.*

Conditions (i) and (ii) and the first equality in (iii) are analogous to the case with no fixed effects in Section 2. The second equality in (iii) holds if (U_i, V_i, η, ζ) is multivariate normal.

As in Section 4.1, plugging in the reduced form of $\overline{Y}$ gives:

$$Y = \tilde{\beta}_0 + X\beta + \overline{X}\tilde{\gamma} + E\left(U^* + \frac{\alpha}{1-\alpha}\overline{U}^* \middle| X, Z, \mathscr{S}\right) + \tilde{\varepsilon}, \tag{11}$$

where $E(\tilde{\varepsilon}|X,Z,\mathscr{S}) = 0$ by construction. We can estimate the model using the following steps:

Step 1. Use a *correlated random effect* model, as proposed in Chamberlain (1980), to estimate the selection equation in Equation (10). Let $F(\zeta|Z)$ denote the distribution of ζ conditional on Z, which is parameterized up to some unknown parameters. For example, following Chamberlain (1980), we may adopt the specification below for the fixed effects in group participation decisions:

Assumption 1' (continued). *(iv) $\zeta = \overline{Z}\tau + e$, where $\overline{Z}$ is the average of individual Z_i's within the group, and $e \perp Z$ with $e \sim N(0, \sigma_e^2)$.*

Under Assumption 1' (iv), the distribution $F(\zeta|Z)$ is parameterized up to (τ, σ_e). We estimate them jointly with δ using maximum likelihood:

$$(\hat{\delta}, \hat{\tau}, \hat{\sigma}_e) = \arg\max_{\delta,\tau,\sigma_e} \sum\nolimits_{g\leq G} \log \int \prod\nolimits_{i\leq n} \tilde{\Phi}_{g,i}(e;\delta,\tau)^{D_{g,i}} \left[1 - \tilde{\Phi}_{g,i}(e;\delta,\tau)\right]^{1-D_{g,i}} \frac{1}{\sigma_e}\phi\left(\frac{e}{\sigma_e}\right) de,$$

where $\tilde{\Phi}_{g,i}(e;\delta,\tau) \equiv \Phi(Z_{g,i}\delta + \overline{Z}_g\tau + e)$. Here subscripts g index the groups in the sample, and we use $D_{g,i}, Z_{g,i}$ to denote variables for i in a potential group g.

Step 2. Apply a generalized method to correct the bias due to sample selection, using estimates for (δ, τ, σ_e) from the previous step. Let $\overline{\mathscr{S}}$ denote the event that "$V_j^* \geq -Z_j\delta$

for all $j \leq n$ and $V_k^* < -Z_k\delta$ for all other $k \in \mathscr{N}$". For an actual group member i,

$$E(U_i^*|X,Z,\mathscr{S}) \;=\; E(U_i^*|\overline{\mathscr{S}}) = \int E(U_i^*|\zeta, V_i \geq -Z_i\delta - \zeta)dF(\zeta|\overline{\mathscr{S}}), \qquad (12)$$

where the second equality is due to Assumption 1' (iii) above. The integrand on the right-hand side of (12) is:

$$\begin{aligned} E(U_i^*|\zeta, V_i \geq -Z_i\delta - \zeta) = &\quad \int E(U_i^*|\zeta, V_i)dF(V_i|V_i \geq -Z_i\delta - \zeta) \\ = &\quad \int (\pi_1\zeta + \pi_2 V_i)\, dF(V_i|V_i \geq -Z_i\delta - \zeta) \\ = &\quad \pi_1\zeta + \pi_2\lambda(Z_i\delta + \zeta). \end{aligned}$$

Under Assumption 1'-(ii), (iii) and (iv), we write the right-hand side of (12) as:

$$\int \left[\pi_1(\overline{Z}\tau + e) + \pi_2\lambda(Z_i\delta + \overline{Z}\tau + e)\right] dF(e|\mathscr{S}^*).$$

where $\mathscr{S}^*$ denotes the event "$e + V_j \geq -Z_j\delta - \overline{Z}\tau$ for all $j \leq n$, and $e + V_k < -Z_k\delta - \overline{Z}\tau$ for all other $k \in \mathscr{N}$". Under (i) and (iv) in Assumption 1', $e = \zeta - \overline{Z}\tau$ is independent from the vector of selection errors V. Then the conditional distribution of

$$e \mid e + V_1 = t_1, e + V_2 = t_2, ...e + V_{(\#\mathscr{N})} = t_{(\#\mathscr{N})}$$

is normal with variance $\tilde{\sigma}^2 \equiv \left[\sigma_e^{-2} + (\#\mathscr{N})\right]^{-1}$ and mean $\tilde{\sigma}^2(\sum_{i \in \mathscr{N}} t_i)$ (where $\#\mathscr{N}$ denotes the cardinality of $\mathscr{N}$). Therefore, we can write (12) in the form of

$$\underbrace{\pi_1 \int (\overline{Z}\tau + e)d\tilde{F}(e|Z;\theta)}_{\psi(Z)} + \underbrace{\pi_2 \int \lambda(Z_i\delta + \overline{Z}\tau + e)d\tilde{F}(e|Z;\theta)}_{\varphi_i(Z)},$$

where $\tilde{F}(e|Z;\theta)$ is the distribution of e conditional on $\mathscr{S}^*$. This conditional distribution is known up to (δ, τ, σ_e), which can be consistently estimated from Step 1. The quantities ψ, φ_i can be constructed using these estimates of (δ, τ, σ_e). Note φ_i varies across individual members in each group while ψ does not.

Step 3. Using the estimates from Steps 1 and 2, we can write the individual outcome Y_i

in (11) as

$$Y_i = \tilde{\beta}_0 + X_i\beta + \overline{X}\tilde{\gamma} + \tilde{\pi}_1\psi + \pi_2\varphi_i + \tilde{\pi}_2\overline{\varphi} + \tilde{\varepsilon}_i,$$

where $\overline{\varphi}$ denotes the group mean of φ_i, and $\tilde{\pi}_1 \equiv \frac{\pi_1}{1-\alpha}$, $\tilde{\pi}_2 \equiv \frac{\alpha\pi_2}{1-\alpha}$. Let $\widetilde{W}_i \equiv (1, X_i, \overline{X}, \psi, \varphi_i, \overline{\varphi})$. Provided $E(\widetilde{W}_i'\widetilde{W}_i|\mathscr{S}^*)$ has a full rank, we can consistently estimate $(\tilde{\beta}_0, \beta', \tilde{\gamma}', \tilde{\pi}_1, \pi_2, \tilde{\pi}_2)$ by regressing Y_i on $\widetilde{W}_i$ in the sample. This in turn allows us to construct consistent estimators for α, β_0 and γ as before.

5 Monte Carlo

We present two Monte Carlo experiments, with different sizes of potential groups, $\#\mathscr{N} = 10$ and $\#\mathscr{N} = 50$. For each group g and member i, let $X_{g,i}$ and $Z_{g,i}$, be two distinctive scalar variables, drawn from the standard normal distribution independently.[6] Let $(U_{g,i}, V_{g,i})$ be drawn from the bivariate normal with mean $(0,0)$, unit variance and covariance σ_{uv}. These error terms are independent across individuals and groups. The sample selection is given by

$$D_{g,i} = 1\{\delta_0 + \delta_1 X_{g,i} + \delta_2 Z_{g,i} + V_{g,i} \geq 0\},\ i \in \mathscr{N}_g, g = 1, \cdots, G,$$

with $\delta = (0,1,1)$. Among the actual group members with $D_{g,i} = 1$, the outcomes are generated through the reduced form:

$$Y_{g,i} = \frac{\beta_0}{1-\alpha} + \beta X_{g,i} + \overline{X}_g \frac{\alpha\beta + \gamma}{1-\alpha} + U_{g,i} + \frac{\alpha}{1-\alpha}\overline{U}_g.$$

We set $(\alpha, \beta_0, \beta, \gamma, \sigma_{uv}) = (1/2, 1, 1, 1, 2/3)$. We experiment with sample sizes $G = 250, 500, 1000, 2000$, and report average biases and mean-squared error (MSE) with 1,000 replications in Tables 1 and 2.

[6] With a slight abuse of notation, in this section, we use $Z_{g,i}$ to denote the instrument variable that enters the selection equation, but not directly in the outcome equation.

Table 1: Monte Carlo Results: #$\mathscr{N}$=10

G	Average Bias				
	α	β_0	β	γ	σ_{uv}
250	-0.040	0.112	0.001	0.141	-0.002
500	-0.015	0.043	0.000	0.055	0.001
1,000	-0.007	0.020	0.000	0.024	-0.001
2,000	-0.006	0.015	0.000	0.021	0.000
	MSE				
	α	β_0	β	γ	σ_{uv}
250	0.042	0.315	0.002	0.557	0.006
500	0.011	0.075	0.001	0.143	0.003
1,000	0.005	0.036	0.000	0.068	0.002
2,000	0.002	0.018	0.000	0.034	0.001

Table 2: Monte Carlo Results: #$\mathscr{N}$=50

G	Average Bias				
	α	β_0	β	γ	σ_{uv}
250	-0.032	0.088	0.000	0.113	0.000
500	-0.014	0.039	0.000	0.054	0.000
1,000	-0.006	0.016	0.000	0.022	0.001
2,000	-0.003	0.009	0.000	0.011	0.000
	MSE				
	α	β_0	β	γ	σ_{uv}
250	0.026	0.199	0.000	0.343	0.001
500	0.010	0.077	0.000	0.135	0.001
1,000	0.004	0.033	0.000	0.056	0.000
2,000	0.002	0.015	0.000	0.026	0.000

In Tables 1 and 2, both the average bias and MSE decrease at the same rate as the sample size increases. This confirms our asymptotic theory that the two-step estimator is root-G consistent. Convergence of the squared average bias at a rate faster than the increase in sample sizes indicates the dominant component in MSE is the estimator variance. Meanwhile, the size of groups does not have an obvious impact on estimation precision, especially in larger samples.

6 Peer Effects in Online Training

In this section, we apply the model to estimate peer effects in a large online teacher training program in China, known as the Young Teacher Empowerment Program (YTEP).[7] The YTEP is an annual training program designed to boost the morale and to improve the skills of young teachers in elementary and middle schools in rural China. To participate in the YTEP program, applicants must be chosen by participating rural schools and the education bureau in the local county government coordinating the training.

Our data was collected from the Training Year of 2019-2020, which consists of two semesters (Fall 2019 and Spring 2020).[8] The YTEP consists of two phases, mandatory general courses in Fall 2019 and elective field courses in Spring 2020. All trainees are automatically enrolled in two mandatory courses, *Career Development* and *Teacher Ethics*. We investigate how peer effects affect the trainees' lecture attendance in these mandatory courses, which have a much larger enrollment than elective field courses.

The sample contains 8,627 trainees across 63 counties in 17 provinces of China. The *Career Development* and *Teacher Ethics* courses consist of 17 and 12 independent, synchronous lectures respectively.[9] Instructors and contents differ across lectures in each course. We collect data from 29 lectures in the sample. For each lecture, trainees first decide whether to attend the lecture, and then decide on the duration of attendance (for how long to stay in the lecture). We define all teachers from a county enrolled in the program as a *potential* group for a specific lecture. Their decisions to attend lectures are modeled as in (1). Potential group members who attended a lecture form an *actual* group, which is county-lecture specific. We model actual group members' duration of lecture

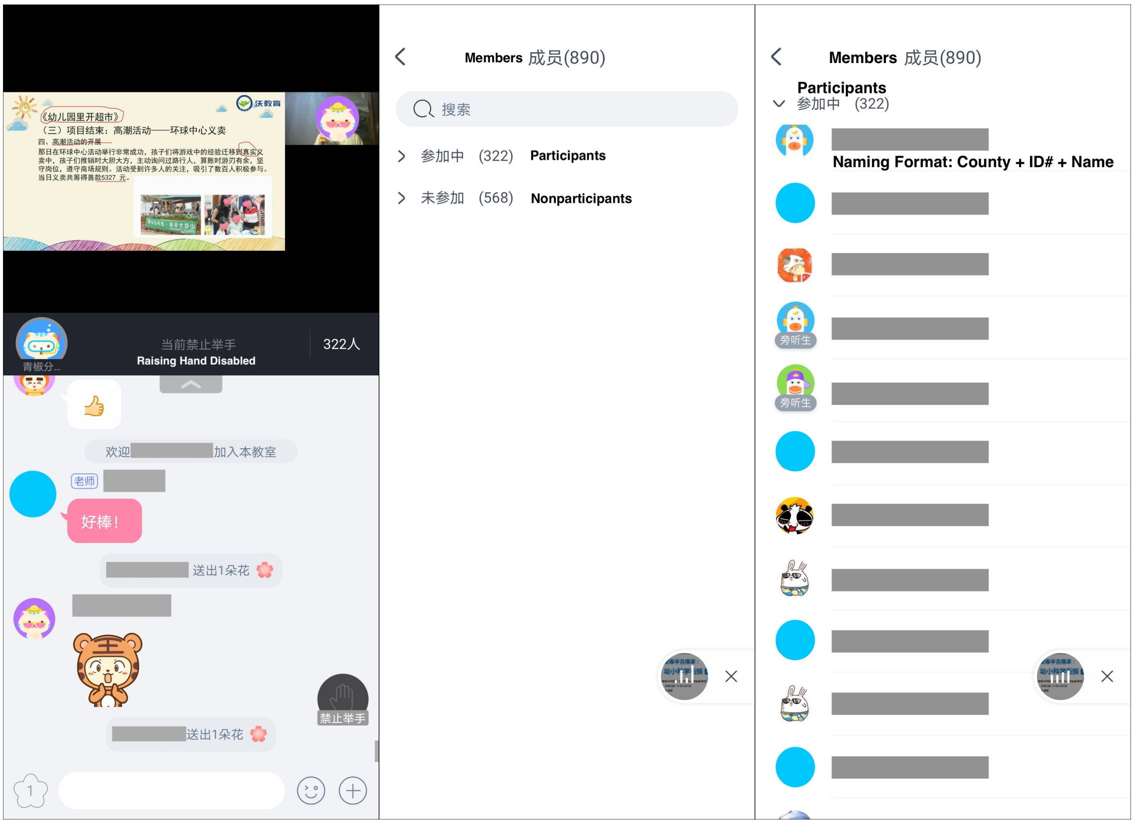
[7] YTEP details: http://www.youcheng.org/news_detail.php?id=645

[8] The data from the YTEP have been used in economics research (Li et al., 2022, Ma et al., 2023).

[9] For the career development course, the first two lectures are the opening ceremony and the program introduction, and the last one is the semester closing ceremony.

attendance, measured by the number of minutes stayed in a lecture, as the outcomes determined in (2).[10]

As explained, potential peer groups are defined by counties, and indexed by g as in Section 2. In the first stage, for each group g, the set of *potential* participants, denoted by $\mathscr{N}_g$, is defined as all trainees from that county who have enrolled in the courses. In the second stage, an *actual* group is formed by those who are from a county and attended a specific lecture. The size of an actual group is the number of participating trainees, denoted by n_g, for a county-lecture pair. Sample selection takes place when individuals decide to participate in a lecture, thus becoming actual group members. In other words, the selection into an actual group occurs within a predetermined potential group.



《幼儿园里开超市》

（三）项目结束：高潮活动——环球中心义卖

四、高潮活动的开展

那日在环球中心活动举行非常成功，孩子们将游戏中的经验迁移到真实义卖中，孩子们推销时大胆大方，主动询问过路行人，算账时游刃有余，坚守岗位，遵守商场规则。活动受到许多人的关注，吸引了数百人积极参与。当日义卖共筹得善款5327 元。

欢迎[redacted]加入本教室

老师 [redacted]

好棒！

[redacted] 送出1朵花

[redacted]

[redacted]送出1朵花

Members 成员(890)

参加中 (322) Participants

未参加 (568) Nonparticipants

Members 成员(890)

Participants
参加中 (322)

[redacted]

Naming Format: County + ID# + Name

[redacted]

Figure 1: Interface of Instructional Platform

[10] The number of minutes for which trainees stayed in the lecture may exceed the actual instruction time if they entered the virtual classroom early or stayed overtime to ask questions.

The lectures are held synchronously online in the evenings on a weekly basis. Figure 1 shows the user interface of the instructional platform. Upon entering the meeting room, a lecture participant can watch the live broadcast of the lecture. Below the presentation window, there is a chatroom where participants can communicate with the instructor, TAs, and other trainees. More importantly, the participants can observe the total number of other participants and nonparticipants (those who enrolled in the course but did not attend this particular lecture) as well as the list of participants. The list tracks entry or exit in real time.

Each trainee has a unique identifier formatted as "*County + ID# + Name*". The list is sorted by the characters of identifiers, and participants from the same county are placed adjacently. Thus, a lecture participant could easily observe peers from the same county sitting in the lecture, and be subject to potential peer effects. In addition, each county has a coordinator who helps with program administration and communicates with the trainees. County coordinators are usually local education administrators. They organize trainees from the same county into a group via online social platforms, such as an online WeChat group.[11] Therefore, all trainees from the same county naturally form a potential peer group for a specific lecture. For trainees attending a lecture, we observe the duration of their attendance.

We obtain information about trainees from their registration records, as well as their responses to routine program surveys during the training period (designed to collect feedback). The program surveys consist of two waves, one at the end of the first semester and the other upon completion of the program at the end of the second semester. Each wave has a response rate of about 40%.The registration and survey provide demographic features about the trainees and characteristics of their school and the county. We impute missing values of the county characteristic, *Encouraging County Coordinator*, based on

[11] WeChat is a popular Chinese instant messaging smartphone application, similar to WhatsApp.

the survey responses available.[12] Table 3 summarizes the related variables (variables with no designated units are dummies).

Table 3: Summary Statistics

Variable	Obs	Mean	SD	Min	Max
Selection Stage					
Lecture Attendance	250,183	0.404	0.491	0	1
Outcome Stage					
Duration of Attendance (mins)	100,954	65.083	24.120	1	165
Personal Characteristics					
Married	5,357	0.295	0.456	0	1
Gender (male)	8,573	0.230	0.421	0	1
Teaching Experience (yrs)	5,357	2.145	3.030	0	37
Teachers College	8,135	0.715	0.451	0	1
Tenured Teacher	8,452	0.452	0.498	0	1
Bachelor's Degree (or higher)	8,233	0.800	0.400	0	1
Ethnicity (Han)	7,667	0.671	0.470	0	1
Slow Internet Speed	5,357	0.200	0.400	0	1
Village School	8,453	0.415	0.493	0	1
County Characteristics					
Encouraging County Coordinator	8,000	0.255	0.436	0	1
Other Statistics					
County Enrollment	63	136.778	157.221	5	1,020
Group Size	1,517	48.238	81.738	2	797

Note: Lecture Attendance and Duration of Attendance are auto-recorded by the instruction platform. Lecture Attendance records all 8,627 trainees enrolled in the program in 29 lectures, yielding 250,183 observations in total at the individual-lecture level. Duration of Attendance records 100,954 observations with positive duration of attendance. Married, Teaching Experience, Slow Internet Speed, and Encouraging County Coordinator are reported in a program survey. Other characteristics are reported in the program registration.

Specifically, *Lecture Attendance* equals to 1 if a trainee decides to join a lecture, 0 otherwise. As we pool 8,627 trainees' attendance in 29 lectures, we obtain 250,183

[12] For *Encouraging County Coordinator*, we use the mode of values self-reported by survey respondents from the same county to impute the missing values for non-respondents from the same county. In total, 3,487 missing values of *Encouraging County Coordinator* are imputed. The county mode replaces the opposite responses so that the variables have the same values for everyone from the same county. Missing values are assigned in the imputation if there are multiple modes.

observations of participants-lecture pairs. The average rate of lecture attendance is 40.4%. In the outcome stage, we only include observations with positive duration of attendance, leading to a total of 100,954 observations. On average, an individual stays in a lecture for about 65 minutes; the average length of a lecture is around 95 minutes.

Among trainees, nearly 30% are married, and 23% are males. Trainees are mostly young teachers with about 2 years of teaching experience on average; about 72% of them graduated from teachers' colleges. Over 45% of the trainees hold tenured positions, and 80% of them obtained a bachelor's or higher degree.

About 20% of the trainees report slow internet speed, which may deter lecture participation. Over 40% of the trainees work at rural village schools. We also include the behavior of county coordinators, defining "encouraging county coordinator" if the majority of the survey respondents from that county report that their coordinator sent lecture reminders. About 25.5% of the trainees report that their coordinators do so.

We apply our multi-step method to estimate the peer effects on the trainees' duration of lecture attendance, while accounting for self-selection in lecture attendance (participation). Our group definition has two overlays, namely county and lecture. In our context, for simplicity, joining a school/county upon employment is assumed to be exogenously given. Self-selection occurs when trainees decide to attend a lecture after enrolling in the training course, governed by (1).

In the first (selection) stage, we model the self-selection in attending a lecture and becoming *actual* group members with the following specification:

$$Attendance_{g,i} = 1\{\delta_0 + Z'_{g,i}\delta + \nu_{g,i} \geq 0\}, \tag{13}$$

where $Attendance_{g,i}$ is 1 if trainee i from the *potential* group g showed up in a lecture, and is 0 otherwise. For the instrumental variables in $Z_{g,i}$ that affect the decision to attend a lecture but not the duration of attendance, we use the marital status, *Married*, and its interaction with other covariates. It is likely that the decision to attend evening lectures generally requires planning and coordinating with spouses. On the other hand, the spouses may have little influence on the duration of lecture attendance, once a teacher already decides and commits to joining the lecture.

We also control for lecture-level fixed effects by including lecture dummies in $Z_{g,i}$.[13] These lecture dummies are meant to capture lecture-level heterogeneities that are not measured in the data, such as content relevance or instructor competence. If not controlled for and left in the error terms, these unobserved, lecture-level fixed effects would also result in selection bias. More importantly, the method in our paper addresses the selection bias due to the individual-level in addition to lecture-level unobserved factors. In other words, even after including these lecture dummies, we still need to apply our method because unobserved *individual-level* factors may also lead to endogenous selection into lecture attendance.

Table A1 reports the first-stage probit results. The signs of coefficient estimates are generally consistent with the expectations. For instance, male teachers are less likely than their female peers to attend a lecture. Graduates from teachers' colleges participate in the program more frequently than those without formal teacher training. Tenured trainees are less active than their untenured colleagues. Slow internet connectivity is a significant disincentive for lecture participation. Working at rural village schools tends to decrease married individuals' attendance but incentivizes unmarried ones to join a lecture with a higher probability. A potential explanation is that unmarried rural teachers may utilize the training program as an opportunity to change their career path. Having an active coordinator increases the participation likelihood of unmarried trainees but not that of married ones. A Wald test shows that *Married* and its interaction with other

[13] Such a specification is not subject to the "incidental parameter" problem, because in our context the number of lectures is small and fixed, relative to the much larger number of county-teacher pairs.

covariates are jointly significant at the 1% level.[14] These results indicate that *Married* and its interaction terms do affect the decision of participation (attending a lecture), as required for the relevance of instrumental variables.

We index *actual* groups by subscripts $\tilde{g}$ (actual groups are defined on the level of county-lecture pairs); we index trainees who self-select into each group $\tilde{g}$ by i. In the second (outcome) stage, we adopt the following specification:

$$Duration_{\tilde{g},i} = \tilde{\beta}_0 + X'_{\tilde{g},i}\beta + \overline{X}'_{\tilde{g}}\tilde{\gamma} + \sigma_{uv}\lambda_{\tilde{g},i} + \tilde{\sigma}_{uv}\bar{\lambda}_{\tilde{g}} + \varepsilon_{\tilde{g},i}, \tag{14}$$

where $Duration_{\tilde{g},i}$ records how long individual i stays in a lecture, and $X_{\tilde{g},i}$ contains individual demographic variables in Table 3, except for *Married*, which is used as the instrument discussed above. The covariates $X_{\tilde{g},i}$ also include a vector of lecture dummies as in the first stage.

Our method applies to samples where the number of groups (county-lecture combinations) is large relative to the size of each group. To allow for peer and contextual effects, we restrict our groups to have at least two trainees from the same county attending the same lecture. This leads to a sample of 1,517 actual groups with an average group size of 48. The 90th percentile of group size is 89, which is still relatively smaller than the number of groups in the data-generating process.

The vector $\overline{X}'_{\tilde{g}}$ consists of the group averages of $X_{\tilde{g},i}$ with each county-lecture combination, as well as the county characteristic. The variable $\lambda_{\tilde{g},i}$ is the inverse Mills ratio constructed from the estimates in Table A1, and $\bar{\lambda}_{\tilde{g}}$ is the group average of $\lambda_{\tilde{g},i}$ within a county-lecture pair. The inclusion of $\lambda_{\tilde{g},i}$ and $\bar{\lambda}_{\tilde{g}}$ helps us to deal with two sources of endogeneity at the same time: self-selection into participation (lecture attendance) and simultaneity in the determination of peer outcomes. Specifically, the reduced form contains a composite error that includes the individual structural error and those of other

[14] The Wald-statistic for testing the joint significance of all interaction terms involving *Married* is 88.8, with 10 degrees of freedom.

group members. To correct for sample selection, $\lambda_{\tilde{g},i}$ takes care of the correlation between V_i and U_i, and $\bar{\lambda}_{\tilde{g}}$ addresses the correlation between V and $\overline{U}$. To reiterate, the inclusion of $\lambda_{\tilde{g},i}$ and $\bar{\lambda}_{\tilde{g}}$ in the reduced form provides variations that enable us to solve the reflection problem.

Table A2 reports the OLS estimates and standard errors. The standard errors are calculated using 1,000 bootstrap replications done by resampling the actual groups with replacement. Generally, individual characteristics have statistically significant effects on the duration of lecture attendance in the reduced form, with signs consistent with intuition. For instance, slow internet speed has a negative impact on attendance duration, which is significant at the 10% level. The statistically significant coefficients of $\lambda_{\tilde{g},i}$ and $\bar{\lambda}_{\tilde{g}}$ indicate a high correlation between the error terms in the selection stage and those in the outcome stage. The large and significant negative coefficient of the inverse Mills ratio is also documented for other contexts in the literature (Heckman, 1976, 1977).

By plugging the estimates of reduced-form parameters from Table A2 into (7), we obtain the estimates of structural parameters, which are reported in Table 4. As noted above, the inclusion of $\lambda_{\tilde{g},i}$ and $\bar{\lambda}_{\tilde{g}}$ in the second-stage regression accounts for the self-selection in lecture attendance. Additionally, with the inclusion of $\lambda_{\tilde{g},i}$ and $\bar{\lambda}_{\tilde{g}}$, the order and rank conditions for recovering the structural coefficients from the reduced-form coefficients are satisfied. The standard errors are estimated using bootstrap resampling.

The estimate of peer effects, i.e., α in (2), is 0.681, and the coefficient is statistically significant at the 1% level. We may interpret peer effects α as the marginal effect of the group average in the structural form (Gaviria and Raphael, 2001, Bramoullé, Djebbari, and Fortin, 2020, Sacerdote, 2011). That is, a 10-minute increase in the average duration of attendance among group peers leads to an 6.81-minute increase in one's own duration of attendance. Our estimate of peer effects is within the unit interval $(0,1)$, which is comparable with those reported for other contexts in the literature (Calvó-Armengol, Patacchini, and Zenou, 2009, Bramoullé, Djebbari, and Fortin, 2009, Lin, 2010).

Table 4: Estimates of Social Effects in Structural Equation

Variable	Estimate	Standard Error
Peer Effects	0.681^{***}	0.104
Intercept	60.633^{***}	13.576
Gender (male)	0.486	0.462
Teaching Experience (yrs)	0.013	0.036
Teachers College	-0.992^{***}	0.236
Tenured Teacher	2.296^{***}	0.413
Bachelor's Degree (or higher)	-0.796^{***}	0.259
Ethnicity (Han)	1.413^{***}	0.249
Slow Internet Speed	-0.496^{*}	0.282
Village School	1.085^{***}	0.219
Average Gender (male)	-2.141	1.349
Average Teaching Experience	-0.215^{***}	0.059
Average Teachers College	0.550	0.476
Average Tenured Teacher	-0.262	0.347
Average Bachelor's Degree (or higher)	1.315^{**}	0.513
Average Ethnicity (Han)	-0.803^{**}	0.345
Average Slow Internet Speed	-1.633	1.271
Average Village School	-0.583^{*}	0.335
Encouraging County Coordinator	-0.434^{***}	0.154

The standard errors are estimated by bootstrap resampling on the groups with replacement for 1,000 replications. Significance Level: *** 1%, ** 5%, * 10%.

The contextual effects of some covariates, such as *Teaching Experience*, *Bachelor's Degree (or higher)*, and *Village School*, are statistically significant in the structural form. One tends to stay shorter in a lecture if the actual group members are more experienced, possibly driven by a large proportion of new teachers who decide to leave the lectures sooner if there aren't enough peers with similar seniority. In some cases, the direct effects (β) and the corresponding contextual effects (γ) have opposite signs, which may depend on whether trainees with certain characteristics are substitutes or complements in the determination of attendance duration (Blume, Brock, Durlauf, and Jayaraman, 2015).

For comparison, we also estimate the model of peer effects in (2) while intentionally ignoring the sample selection issue in (1). However, in this case, without exogenous variation from instruments in the selection equation, we need other exclusion restrictions to solve the reflection problem.[15] In particular, we exploit a variable (internet speed) that has a direct effect on an individual's own duration of attendance, but has no contextual effect on others' in the structural form. Such an exclusion restriction is known to help solve the reflection problem in social interactions models (Manski, 1993, Moffitt, 2001).

We posit that internet speed has a direct effect on an individual's own outcome (duration of lecture attendance), but no contextual effect on others' outcomes. The latter means an individual's duration of attendance is not immediately affected by the proportion of peers who have slow internet access. This is confirmed by results in Table 4, which show the proportion of group peers with slow internet has no significant effect on an individual's duration of attendance. Hence, to estimate (2) while ignoring sample selection due to (1), we exploit this exclusion restriction on the internet speed. [16]

We now provide details in how to estimate (2) while ignoring sample selection and using internet speed as a non-contextual variable. Let I denote the vector of dummy variables indicating slow internet speed for each member in a group, and let X denote the vector of all other covariates. The structural form is

$$Y = \alpha\overline{Y} + \beta_0 + X'\beta + \overline{X}'\gamma + \beta_I I + U, \tag{15}$$

which implies the following reduced form:

$$Y = \tilde{\beta}_0 + X'\beta + \overline{X}'\tilde{\gamma} + \beta_I I + \tilde{\gamma}_I \bar{I} + \widetilde{U}, \tag{16}$$

[15] We have $2K+2$ parameters in the structural form (2). Neglecting sample selection would lead to a reduced form that drops λ_i and $\bar{\lambda}$ from (6) and only has $2K+1$ parameters. Hence, the order condition for recovering structural parameters does not hold.

[16] Apart from internet speed, two other variables (*Teachers College* and *Tenured Teacher*) also show significant direct effects but no statistically significant contextual effects in Table 4. We did not consider them as candidates satisfying exclusion restrictions, because the literature has documented evidence of education contextual effects (Harmon, Fisman, and Kamenica, 2019, Laliberté, 2021), and tenure-related peer effects (De Grip, Sauermann, and Sieben, 2016).

where $\widetilde{U}, \tilde{\beta}_0, \tilde{\gamma}$ are defined as in (3) and $\tilde{\gamma}_I = \frac{\alpha\beta_I}{1-\alpha}$. With the covariate support satisfying the rank condition for the OLS regression in (16), we identify $\tilde{\beta}_0, \beta, \tilde{\gamma}, \beta_I, \tilde{\gamma}_I$, and then recover all structural parameters in (15) as

$$\alpha = \frac{\tilde{\gamma}_I}{\beta_I + \tilde{\gamma}_I};\ \beta_0 = (1-\alpha)\tilde{\beta}_0;\ \gamma = (1-\alpha)\tilde{\gamma} - \alpha\beta. \tag{17}$$

For estimation, we regress $Y_{g,i}$ on $X_{g,i}, \overline{X}_g, I_{g,i}, \overline{I}_g$ and an intercept, using the same sample for the model with sample selection, and then use (17) to calculate the estimates for structural parameters.

Table A3 reports the estimates of the reduced form in (16); Table 5 reports estimates of structural parameters using (17). To illustrate the value of our method in Section 2, we use the model underlying Table 5 (i.e., ignoring sample selection, and using internet speed as an excluded variable with no contextual effects to solve the reflection problem) as a close comparison to our selection model.[17]

Some estimates for direct and contextual effects in these two tables are comparable in terms of signs and significance, despite different identification and estimation strategies. Nonetheless, Table 5 shows that ignoring the sample selection in lecture attendance, we obtain an estimate of 0.838 for the peer effect, which is statistically significant at the 1% level. This estimate differs from that reported in Table 4 (0.681), and the discrepancy is large relative to the standard errors reported.

The distinction between the peer effect estimates in Table 4 and Table 5 illustrates the consequence of failing to account for sample selection bias due to endogenous lecture participation. Specifically, ignoring the sample selection has led to overestimation of peer effects in this context. Intuitively, teachers decide to stay in lectures for longer because of their latent (unobservable) self-motivation, as well as positive peer effects.

[17] To reiterate, the model in Table 4 accommodates endogenous self-selection into lecture participation, and constructs correction terms in the first stage to resolve the reflection (identification) problem. In comparison, the model in Table 5 relies on an exclusion restriction, which is supported by estimates from the selection model, to solve the identification issue.

Not accounting for sample selection in estimation would mean neglecting the role of such latent motivation factors, and instead attributing their impact on lecture attendance to the structural peer effects. It is therefore not surprising that ignoring sample selection has led to overstating these peer effects in our setting. On the other hand, once sample selection is properly accounted for, the peer effects turn out to be more moderate, as is confirmed in our analysis above.

Table 5: Structural Estimates Ignoring Sample Selection

Variable	Estimate	Standard Error
Peer Effects	0.838***	0.052
Intercept	15.254***	4.846
Gender (male)	−1.765***	0.235
Teaching Experience (yrs)	0.085**	0.036
Teachers College	−0.089	0.194
Tenured Teacher	0.316	0.285
Bachelor's Degree (or higher)	−0.391	0.245
Ethnicity (Han)	1.154***	0.234
Slow Internet Speed	−1.427***	0.231
Village School	1.069***	0.220
Average Gender (male)	−0.134	0.772
Average Teaching Experience	−0.149***	0.042
Average Teachers College	0.281	0.272
Average Tenured Teacher	−0.250	0.304
Average Bachelor's Degree (or higher)	0.957**	0.378
Average Ethnicity (Han)	−0.993***	0.282
Average Slow Internet Speed	–	–
Average Village School	−0.794***	0.254
Encouraging County Coordinator	−0.100	0.067

The standard errors are estimated by bootstrap resampling on the groups with replacement for 1,000 replications. Significance Level: *** 1%, ** 5%, * 10%.

7 Conclusion

This paper estimates peer effects in self-selected groups that are formed out of endogenous individual participation decisions. We correct the sample selection bias using individual instruments that affect the participation decisions, but do not directly affect the outcomes. In the context of social interactions, dealing with sample selection requires the insertion of both individual-level and group-level composite correction terms. The inclusion of this latter, group-level correction term provides additional sources of exogenous variation that help us to resolve the reflection problem.

We apply our method to study peer effects in an online teacher training program in China, where the trainees endogenously decide to participate in online lectures. We find significant peer effects in the duration of lecture attendance among trainees after accounting for sample selection bias. Our analysis also shows ignoring the sample selection in this context would result in overestimation of peer effects.

References

Auerbach, E. (2022). Identification and estimation of a partially linear regression model using network data. *Econometrica 90*(1), 347–365.

Badev, A. (2021). Nash equilibria on (un) stable networks. *Econometrica 89*(3), 1179–1206.

Blume, L. E., W. A. Brock, S. N. Durlauf, and R. Jayaraman (2015). Linear social interactions models. *Journal of Political Economy 123*(2), 444–496.

Boucher, V. (2016). Conformism and self-selection in social networks. *Journal of Public Economics 136*, 30–44.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009). Identification of peer effects through social networks. *Journal of Econometrics 150*(1), 41–55.

Bramoullé, Y., H. Djebbari, and B. Fortin (2020). Peer effects in networks: A survey. *Annual Review of Economics 12*, 603–629.

Brock, W. A. and S. N. Durlauf (2001a). Discrete choice with social interactions. *The Review of Economic Studies 68*(2), 235–260.

Brock, W. A. and S. N. Durlauf (2001b). Interactions-based models. In *Handbook of econometrics*, Volume 5, pp. 3297–3380. Elsevier.

Calvó-Armengol, A., E. Patacchini, and Y. Zenou (2009). Peer effects and social networks in education. *The Review of Economic Studies 76*(4), 1239–1267.

Carrell, S. E., B. I. Sacerdote, and J. E. West (2013). From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica 81*(3), 855–882.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *The Review of Economic Studies 47*(1), 225–238.

Dahl, G. B., K. V. Løken, and M. Mogstad (2014). Peer effects in program participation. *American Economic Review 104*(7), 2049–74.

De Giorgi, G., M. Pellizzari, and S. Redaelli (2010). Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics 2*(2), 241–75.

De Grip, A., J. Sauermann, and I. Sieben (2016). The role of peers in estimating tenure-performance profiles: Evidence from personnel data. *Journal of Economic Behavior & Organization 126*, 39–54.

Gaviria, A. and S. Raphael (2001). School-based peer effects and juvenile behavior. *Review of Economics and Statistics 83*(2), 257–268.

Glaeser, E. L., B. Sacerdote, and J. A. Scheinkman (1996). Crime and social interactions. *The Quarterly Journal of Economics 111*(2), 507–548.

Goldsmith-Pinkham, P. and G. W. Imbens (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics 31*(3), 253–264.

Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica 76*(3), 643–660.

Gronau, R. (1974). Wage comparisons–a selectivity bias. *Journal of Political Economy 82*(6), 1119–1143.

Harmon, N., R. Fisman, and E. Kamenica (2019). Peer effects in legislative voting. *American Economic Journal: Applied Economics 11*(4), 156–80.

Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica 42*(4), 679–694.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pp. 475–492. NBER.

Heckman, J. J. (1977). Sample selection bias as a specification error (with an application to the estimation of labor supply functions). *NBER Working paper (No. w0172).*

Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica 46*(4), 931–959.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.

Hong, H., J. D. Kubik, and J. C. Stein (2004). Social interaction and stock-market participation. *The Journal of Finance 59*(1), 137–163.

Hoshino, T. (2019). Two-step estimation of incomplete information social interaction models with sample selection. *Journal of Business & Economic Statistics 37*(4), 598–612.

Hoxby, C. M. (2000). Peer effects in the classroom: Learning from gender and race variation.

Hsieh, C.-S. and L. F. Lee (2016). A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics 31*(2), 301–319.

Jochmans, K. (2023). Peer effects and endogenous social interactions. *Journal of Econometrics 235*(2), 1203–1214.

Johnsson, I. and H. R. Moon (2021). Estimation of peer effects in endogenous social networks: Control function approach. *The Review of Economics and Statistics 103*(2), 328–345.

Laliberté, J.-W. (2021). Long-term contextual effects in education: Schools and neighborhoods. *American Economic Journal: Economic Policy 13*(2), 336–77.

Lee, L.-F. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics 140*(2), 333–374.

Li, H., M. Ma, and Q. Liu (2022). How the covid-19 pandemic affects job sentiments of rural teachers. *China Economic Review 72*, 101759.

Lin, X. (2010). Identifying peer effects in student academic achievement by spatial autoregressive models with group unobservables. *Journal of Labor Economics 28*(4), 825–860.

Lin, Z. and X. Tang (2022). Solving the reflection problem in social interactions models with endogeneity. *Working paper*.

Ma, M., Q. Liu, and H. Li (2023). Program design, personality traits, and online learning participation. *Working paper*.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies 60*(3), 531–542.

Manski, C. F. (2000). Economic analysis of social interactions. *Journal of Economic Perspectives 14*(3), 115–136.

Moffitt, R. A. (2001). Policy interventions, low-level equilibria, and social interactions. *Social Dynamics 4*(45-82), 6–17.

Mora, T. and J. Gil (2013). Peer effects in adolescent bmi: evidence from spain. *Health Economics 22*(5), 501–516.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2111–2245.

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics 116*(2), 681–704.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the Economics of Education*, Volume 3, pp. 249–277. Elsevier.

Sheng, S. and X. Sun (2021). Identification and estimation of social interactions in endogenous peer groups. *Working paper, UCLA and Simon Frasier U*.

Topa, G. (2001). Social interactions, local spillovers and unemployment. *The Review of Economic Studies 68*(2), 261–295.

Trogdon, J. G., J. Nonnemaker, and J. Pais (2008). Peer effects in adolescent overweight. *Journal of Health Economics 27*(5), 1388–1399.

Appendix

Table A1: Selection Stage Probit Estimation Results

Variable	Estimate	Standard Error
Intercept	-0.727^{***}	0.025
Gender (male)	-0.168^{***}	0.010
Teaching Experience (yrs)	-0.004	0.003
Teachers College	0.050^{***}	0.009
Tenured Teacher	-0.155^{***}	0.009
Bachelor's Degree (or higher)	0.050^{***}	0.011
Ethnicity (Han)	-0.021^{**}	0.009
Slow Internet Speed	-0.063^{***}	0.010
Village School	0.021^{**}	0.009
Encouraging County Coordinator	0.049^{***}	0.010
Teaching Experience Squared	0.000	0.000
Married	0.000	0.029
Gender (male) × Married	-0.013	0.020
Teaching Experience × Married	0.009^{***}	0.003
Teachers College × Married	0.066^{***}	0.017
Tenured Teacher × Married	0.027^{*}	0.016
Bachelor's Degree (or higher) × Married	-0.055^{***}	0.020
Ethnicity (Han) × Married	0.012	0.017
Slow Internet Speed × Married	-0.017	0.020
Village School × Married	-0.073^{***}	0.016
Encouraging County Coordinator × Married	-0.090^{***}	0.017
Lecture Fixed Effects	Yes	

Wald Test for joint significance of IVs (Married and its interactions): χ^2 = 88.8, d.f.=10, p-value <0.001.

Significance Level: *** 1%, ** 5%, * 10%.

Table A2: Regression Results in Outcome Stage
(Dependent variable: duration of lecture attendance in minutes)

Variable	Estimate	Standard Error
Intercept	190.067^{***}	20.368
Gender (male)	0.486	0.462
Teaching Experience (yrs)	0.013	0.036
Teachers College	-0.992^{***}	0.236
Tenured Teacher	2.296^{***}	0.413
Bachelor's Degree (or higher)	-0.796^{***}	0.259
Ethnicity (Han)	1.413^{***}	0.249
Slow Internet Speed	-0.496^{*}	0.282
Village School	1.085^{***}	0.219
Average Gender (male)	-5.673^{***}	2.195
Average Teaching Experience	-0.646^{***}	0.104
Average Teachers College	-0.394	1.350
Average Tenured Teacher	4.079^{***}	1.376
Average Bachelor's Degree (or higher)	2.423^{**}	1.123
Average Ethnicity (Han)	0.500	0.701
Average Slow Internet Speed	-6.179^{***}	2.025
Average Village School	0.486	0.585
Encouraging County Coordinator	-1.362^{***}	0.361
$\hat{\lambda}$	-23.118^{***}	4.094
$\hat{\hat{\lambda}}$	-49.350^{***}	15.427
Lecture Fixed Effects	Yes	

Note: the standard errors are estimated by bootstrap resampling on the groups with replacement for 1,000 replications. Significance Level: *** 1%, ** 5%, * 10%.

Table A3: Regression in Outcome Stage (Ignoring Sample Selection)

Variable	Estimate	Standard Error
Intercept	94.144***	2.955
Gender (male)	−1.765***	0.235
Teaching Experience (yrs)	0.085**	0.036
Teachers College	−0.089	0.194
Tenured Teacher	0.316	0.285
Bachelor's Degree (or higher)	−0.391	0.245
Ethnicity (Han)	1.154***	0.234
Slow Internet Speed	−1.427***	0.231
Village School	1.069***	0.220
Average Gender (male)	−9.954***	1.921
Average Teaching Experience	−0.478***	0.086
Average Teachers College	1.275	1.235
Average Tenured Teacher	0.088	0.614
Average Bachelor's Degree (or higher)	3.880***	1.193
Average Ethnicity (Han)	−0.159	0.695
Average Slow Internet Speed	−7.383***	2.004
Average Village School	0.629	0.613
Encouraging County Coordinator	−0.620*	0.341
Lecture Fixed Effects	Yes	

Significance Level: *** 1%, ** 5%, * 10%.