Measuring Social Interaction Effects when Instruments are Weak

Stephen L. Ross
Zhentao Shi

Working Paper

2016-033

12/2016

HUMAN CAPITAL AND
ECONOMIC OPPORTUNITY
GLOBAL WORKING GROUP

The University of Chicago
1126 E. 59th Street Box 107
Chicago IL 60637

www.hceconomics.org

Measuring Social Interaction Effects when Instruments are Weak

Stephen L. Ross[*] and Zhentao Shi[**]

[*]Department of Economics, University of Connecticut

[**]Department of Economics, the Chinese University of Hong Kong

Key words: academic performance, hypothesis testing, endogenous peer effects, random assignment, weak instruments

JEL code: C26, C51, I23, J00

Stephen L. Ross, `stephen.l.ross@uconn.edu`, 341 Mansfield Road, University of Connecticut, Storrs, CT 06269-1063, USA. Zhentao Shi: `zhentao.shi@cuhk.edu.hk`, 912 Esther Lee Building, the Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR, China. Zhentao Shi acknowledges support from the Direct Grant for Research No.4052081 at the Chinese University of Hong Kong.

Abstract

Studies that can distinguish between exogenous and endogenous peer effects of social interactions are relatively rare. One recent identification strategy exploits partial overlapping groups of peers. If a student has two groups of separated peers, the peer choices are correlated through that specific student's choice, but one group's attributes are assumed to directly influence neither the other peer group's attributes nor the choices. In the context of academic performance in higher education, however, the evidence of peer effects on academic outcomes has been mixed, creating a potential for weak instruments.

We utilize a period of transition when students were being reassigned to dormitories from a new campus to an old campus. Many groups of roommates were broken up at the end of freshman year, and then combined with other groups of students from the same school in the sophomore year. We find reduced-form evidence that information about a student's previous year roommates can explain the current test scores of their new roommates. However, due to weak instruments, the estimated endogenous effects appear unreasonably large. We draw on weak-IV robust tests, namely the Anderson-Rubin-type S-test (Stock and Wright, 2000) and Kleibergen's Lagrangian multiplier test (Kleibergen, 2005), to provide properly-sized tests for the endogenous effects between the test scores of current roommates and to calculate lower bounds of such effects. These tests strongly reject the null hypothesis of no endogenous effects.

1 Introduction

A large and growing literature documents the causal effects of exposure to peers on academic outcomes in primary schools, secondary schools, and universities using random or quasi-random assignment that arises naturally or administratively. In primary and secondary schools, causal effects are often identified by comparing students who select into the same school, but happen to belong to different cohorts within the school due to minor age differences (Gould, Lavy, and Daniele Paserman, 2009; Lavy and Schlosser, 2011; Lavy, Paserman, and Schlosser, 2012; Friesen and Krauth, 2011; Bifulco, Fletcher, and Ross, 2011). At universities, causal effects of peers have mostly been identified by examining the effects of random dormitory assignments (Sacerdote, 2001; Marmaros and Sacerdote, 2002; Stinebrickner and Stinebrickner, 2006; Foster, 2006) or, in a few cases, random group or classroom assignments (Carrell, Fullerton, and West, 2009; De Giorgi, Pellizzari, and Redaelli, 2009; De Giorgi, Pellizzari, and Woolston, 2012).[1]

On the other hand, studies that are capable of separating between exogenous and endogenous peer effects are much rarer. Such works have had to grapple with the long standing difficulty: any exogenous peer attribute that can instrument for endogenous behaviors might also affect outcomes directly. In the literature of peer effects, this violation of the exclusion restriction is often referred to as the *reflection problem* (Manski, 1993; Brock and Durlauf, 2001; Moffitt, 2001). Endogenous effects can lead to the clustering of outcomes or behaviors above and beyond the clustering that might have been expected based on individuals' observables. Examples that either explicitly or implicitly test for such clustering include studies of crime (Glaeser, Sacerdote, and Scheinkman, 2003; Billings, Deming, and Ross, 2016), employment (Topa, 2001; Bayer, Ross, and Topa, 2008), welfare usage (Bertrand, Luttmer, and Mullainathan, 2000), prenatal care (Aizer and Currie, 2004), youth health behaviors (Weinberg, 2007; Fletcher and Ross, 2012). Angrist (2014) has recently provided a general critique of attempts to detect social interactions using tests for social multipliers

[1] See Ross (2011), Durlauf (2004) and Ioannides and Datcher Loury (2004) for relevant literature reviews.

of this type.[2]

One recent study has directly isolated endogenous or social interaction effects by exploiting partially overlapping groups of peers. De Giorgi, Pellizzari, and Redaelli (2010) use classroom assignment to test whether peer's choice of economic major affects whether students decide to major in Economics. Instruments for the peer choice come from the attributes of a peer's peer from an earlier classroom assignment. Consider, for instance, a random assignment or quasi-random circumstance where a person A is exposed to his peer B at time 2 and B is exposure to his peer C at time 1, but A is never directly exposed to C. Exploiting random assignment at time 1, one can only uncover the reduced-form effect of peer C on peer B's behavior, which is caused by the effect of either peer behaviors or exogenous peer attributes. However, since A is not exposed to C, C can only affect A through the behavior of A's peer B, and as a result the attributes of C can act as instruments for peer B's behavior. Note that we do not require that A and C never meet or interact. Rather, conditional on belonging to our sample and the observed information, A and C are no more likely to have interactions than any other two randomly chosen individuals from our sample, and it is only their joint connection to B that creates detectable relationships between the outcomes of A and C.[3]

We use data from a period of transition. A university opened a new campus to accommodate incoming freshmen, while they relocated these students to the original campus in the next year. This dormitory reassignment broke up many groups of roommates at the end of freshman year, and they were combined with other groups of students in the sophomore year. As a result, we observe three cohorts of students with different roommate assignments in their freshman and sophomore years. Balancing tests suggest roommate assignment is quasi-random, since the freshman test scores of a student cannot be explained by the attributes of the freshman year roommates of that student's current roommates. Given the evidence

[2] See Brock and Durlauf (2001) for more structural approaches to addressing the reflection problem.

[3] For example, if B is studying (or partying) a lot, he may include both A and C in a study group (or in several wild nights out) and in this way A's incidental exposure to C is part of the behavioral spillovers caused by residing with or sharing a class with B.

of quasi-random assignment in this dataset, we exploit De Giorgi, Pellizzari, and Redaelli (2010)'s partially overlapping peer group strategy. Specifically, we estimate the endogenous effect of each student's sophomore test scores on their roommate's sophomore test scores, using that student's freshman year roommates' attributes as instruments for the student's sophomore test score.

However, when considering academic performance in higher education, another problem raises its head in that much of the evidence of strong peer effects in higher education have been associated with effects on social, rather than academic, outcomes. If the effects of peers on academic outcomes are modest in size, then the peer-based instruments may at best weakly identify the resulting model. Evidence of peer effects on academic performance in college is relatively limited. Neither Stinebrickner and Stinebrickner (2006) nor Foster (2006) shows evidence that randomly assigned dormitory peers affect academic performance. Sacerdote (2001) finds that dormitory assignment affects academic effort and membership in social organizations. Marmaros and Sacerdote (2006) detect effects on friendship formation, but their evidence of effects on student grade point average (GPA) in these samples is much more mixed. Finally, as mentioned earlier, De Giorgi, Pellizzari, and Redaelli (2010) provide evidence of only for effects on college major.

In this paper, we examine the effects of quasi-randomly assigned roommates on the academic performance of college students in the School of Economics at a major Chinese university. A reduced-form regression indicates that information about a student's previous year roommates can explain the current test scores of their new roommates. Nevertheless, when using the conventional instrumental variable (IV) method to directly estimate the endogenous effects, we are surprised that the point estimate appears even larger in magnitude than the estimate on the student's own lagged test score. Standard test statistics from our first-stage regression uncovers the problem of weak instrumentation (Staiger and Stock, 1997; Stock and Yogo, 2005; Stock, Wright, and Yogo, 2014).

In view of the unreliability of the conventional IV method in this context, we draw on

Stock and Wright (2000) and Kleibergen (2005) for a properly sized robust test for the endogenous effects between the test scores of current roommates. We further calculate the lower bounds of the confidence intervals for the size of such effects. The tests and the associated lower bounds are robust to any clustering within the data. These tests strongly reject the null hypothesis of no endogenous effects, providing strong and relatively unique quasi-experimental evidence of endogenous peer effects on academic outcomes. Significantly, the lower bound of our 95% confidence interval from the weak-IV robust estimates lies significantly below the counterpart from the traditional instrumental variable estimates. Therefore, our paper also provides a valuable tool for future studies that examine endogenous peer effects.

The rest of the paper is organized as follows. Section 2 specifies the empirical model with endogenous effects, and then briefly discusses how to test those effects under weak instruments. Section 3 introduces the dataset, and Section 4 presents the empirical analysis. Given that empirical evidence supports the maintained assumption of random assignment and suggests the presence of weak IV, we conduct the weak-IV robust tests and elaborate the results based on those tests. Section 5 provides brief concluding remarks.

Notation: Throughout this paper, we will follow the convention in econometrics to denote a plain letter, say x, as a scalar, a lowercase bold $\mathbf{x}$ as a column vector, and an uppercase bold $\mathbf{X}$ as a matrix. We denote $\mathbf{I}_n$ as an $n \times n$ identity matrix, and $\mathbf{1}_n$ as an $n \times 1$ vector of ones.

2 Framework of Empirical Study

2.1 Model of Roommate Peer Effect

Consider a situation where individuals are randomly assigned to dormitory rooms in the first period, and then the same individuals are randomly re-assigned into dormitory rooms in a second period. This reassignment creates situations where an individual i is exposed to new roommates j in the second year, i.e. roommates who have different year 1 roommates than individual i. Those first period exposures that are unique to roommate j can serve as instruments for j's outcomes during the second period. Specifically, they can explain outcomes for individual i only through endogenous effects. By construction, individual i has not been exposed to those circumstances, allowing us to separate the peer effects caused by roommate j's behaviors from the peer effects associated with j's attributes. In order to see how this would work, we provide a simple example using the standard linear-in-means peer effects model.

Example 1. Figure 1 illustrates the formation of a room of three people, index them by $i = 1, 2, 3$, in year 2. In year 1 people 1, 3, 4 and 5 lived in one dormitory room, and people 2, 6 and 7 lived in another room. In year 2 both rooms were split, and people 1, 2, and 3 were new roommates.

[Figure 1 about here.]

We check the influence of individual 1, referred to as the *treatment individual*, on the second period outcome of individual 2, referred to as the *treated individual*. Individual 1 has lived with individuals 3, 4 and 5 in year 1. Since individual 1 resided with individual 3 in both years, while with 4 and 5 only in year 1, we define individuals 4 and 5 as *instrumental individuals* in that their attributes will serve as instruments for the outcome of treatment individual 1 when explaining the outcome of the treated individual 2.

Next, we assume that year 2 outcomes are a linear function of own first period test score, own attributes, average second period peer outcomes, and average second and first period peer attributes where attributes are invariant over time and peer attributes are determined by roommate assignment. Suppose that we observe $\mathbf{x}_i$, a time invariant m-dimensional vector of exogenous attributes, and y_{it}, a scalar outcome in period t. If person 2 lived with individuals 6 and 7 only in year 1, then the year 2 outcome equation for this example is

$$y_{22} = \alpha_1 \frac{y_{12} + y_{32}}{2} + \alpha_2 y_{21} + \boldsymbol{\alpha}_3' \mathbf{x}_2 + \boldsymbol{\alpha}_4' \frac{\mathbf{x}_1 + \mathbf{x}_3}{2} + \boldsymbol{\alpha}_5' \frac{\mathbf{x}_6 + \mathbf{x}_7}{2} + \alpha_6 + \varepsilon_{22}. \tag{1}$$

The first random variable on the right-hand side, $y_{12} + y_{32}$, is endogenous and must be dealt with by an instrumental equation. In order to obtain this instrumental equation, we add together the two equations for y_{12} and y_{32} that follow the same structure as (1):

$$\begin{aligned} y_{12} + y_{32} = {} & \beta_1 (y_{11} + y_{31}) + \boldsymbol{\beta}_2' (\mathbf{x}_1 + \mathbf{x}_3) + \boldsymbol{\beta}_3' \left(\frac{\mathbf{x}_2 + \mathbf{x}_3}{2} + \frac{\mathbf{x}_1 + \mathbf{x}_2}{2} \right) \\ & + \boldsymbol{\beta}_4' \left(\frac{\mathbf{x}_3 + \mathbf{x}_4 + \mathbf{x}_5}{3} + \frac{\mathbf{x}_1 + \mathbf{x}_4 + \mathbf{x}_5}{3} \right) + 2\beta_5 + (\varepsilon_{12} + \varepsilon_{32}) \end{aligned} \tag{2}$$

Notice that individuals 1 and 3 have the same year 1 roommates except that the roommate contemporaneous test scores are dropped from the equation, so that the coefficients change accordingly. We simplify the above equation by collecting terms involving $\mathbf{x}_1$ and $\mathbf{x}_3$:

$$\begin{aligned} y_{12} + y_{32} = {} & \beta_1 (y_{11} + y_{31}) + \boldsymbol{\beta}_2' (\mathbf{x}_1 + \mathbf{x}_3) + \left(\boldsymbol{\beta}_2 + \frac{\boldsymbol{\beta}_3}{2} + \frac{\boldsymbol{\beta}_4}{3} \right)' (\mathbf{x}_1 + \mathbf{x}_3) + \boldsymbol{\beta}_3' \mathbf{x}_2 \\ & + \boldsymbol{\beta}_4' \frac{2}{3} (\mathbf{x}_4 + \mathbf{x}_5) + 2\beta_5 + (\varepsilon_{12} + \varepsilon_{32}) \end{aligned} \tag{3}$$

Equation (3) illustrates that the attributes of the instrumental individuals 4 and 5 explain the second period peer outcomes for the treatment individuals in (2). These variables can be excluded from equation (1), which describes the year 2 outcome for the treated individual. □

More formally and generally, we consider all endogenous effects between the r_{kt} roommates in room k at time t simultaneously. We express $\underset{(r_{kt} \times 1)}{\mathbf{y}_{kt}}$, the vector of period t roommate test scores, as the dependent variable in a linear simultaneous equations system involving $\mathbf{y}_{kt}^{t-1}$, the own lagged test score, $\underset{(r_{kt} \times m)}{\mathbf{X}_{kt}}$, the own attributes, $\mathbf{X}_{kt}^{-i}$, the period t average roommate attributes, and $\mathbf{X}_{kt-1}^{-i}$, the period $t-1$ average roommate attributes.[4] Drawing on (1), we write

$$\mathbf{A}_1\mathbf{y}_{kt} = \mathbf{y}_{kt} - \alpha_1\mathbf{y}_{kt}^{-i} = \alpha_2\mathbf{y}_{kt-1} + \mathbf{X}_{kt}\boldsymbol{\alpha}_3 + \mathbf{X}_{kt}^{-i}\boldsymbol{\alpha}_4 + \mathbf{X}_{kt-1}^{-i}\boldsymbol{\alpha}_5 + \alpha_6\mathbf{I}_{r_{kt}} + \boldsymbol{\varepsilon}_{kt}, \tag{4}$$

where $\mathbf{y}_{kt}^{-i}$ is the vector of average peer scores, and

$$\underset{(r_{kt}\times r_{kt})}{\mathbf{A}_1} = \mathbf{I}_{r_{kt}} - \alpha_1\left(r_{kt}-1\right)^{-1}\left(\mathbf{1}_{r_{kt}}\mathbf{1}_{r_{kt}}' - \mathbf{I}_{r_{kt}}\right)$$

captures the interactions of the current roommate test score—the endogenous effects. The first term on the rightmost side of (4) is the direct effect of the individual student's lagged test score. The next two terms concerning $\mathbf{X}_{kt}$ and $\mathbf{X}_{kt}^{-i}$ capture the direct effect of the individuals own attributes and the peer effects associated with current roommate attributes, respectively. The fourth term involving $\mathbf{X}_{kt-1}^{-i}$ represents the effect of each individual student's period $t-1$ roommate average attributes. The last two terms are simply the intercept and unobservable for each room member's test score. We view $\mathbf{y}_{kt}^{t-1}$, $\mathbf{X}_{kt}$, $\mathbf{X}_{kt}^{-i}$ and $\mathbf{X}_{kt-1}^{-i}$ all as exogenous variables.

One source of the reflection problem arises as $\mathbf{X}_{kt}^{-i}$ and $\mathbf{y}_{kt}$ are from the same group of people. In contrast, $\mathbf{X}_{kt-1}^{-i}$ carries new information outside of that group. As long as a person, say i, in that group lived with a new roommate j, the variation in j's previous roommate attributes serves as information that is peculiar to explain j's period t test score. Therefore, j, a treatment individual for i, can be instrumented by the average attributes of his previous roommates that i has never lived with (i's instrumental individuals). We denote the average attributes of i's instrumental individuals as $\mathbf{z}_{ikt}$, where the subscript refers to the treated individual i in room k at period t. The treated individuals form a restricted subsample of students. Even though $\mathbf{z}_{ikt}$ is part of $\mathbf{X}_{kt-1}^{-i}$, every individual has $\mathbf{X}_{kt-1}^{-i}$ but only the treated

[4] Although $\mathbf{x}_t$ is time-invariant, the subscripts in $\mathbf{X}_{kt}$, $\mathbf{X}_{kt}^{-i}$ and $\mathbf{X}_{kt-1}^{-i}$ identify the room. The superscript "$-i$" in $\mathbf{X}_{kt}^{-i}$ and $\mathbf{X}_{kt-1}^{-i}$ indicates "peers".

individuals has an associated $\mathbf{z}_{ikt}$.

Since we will focus on year 2 in our dataset, we drop the subscript t for concise notation. For each individual i in room k, define a moment function

$$\mathbf{g}_{ik}\left(\boldsymbol{\alpha}\right) = \mathbf{z}_{ik}\left(y_{ik} - \boldsymbol{\alpha}'\mathbf{x}_{ik}\right),$$

where $y_{ik} = y_{ikt}$ is the dependent variable,

$$\underset{((4m+2)\times 1)}{\mathbf{z}_{ik}} = \left(\underset{(1\times 1)}{y_{ikt}^{t-1}}, \underset{(1\times m)}{\mathbf{x}_{ikt}'}, \underset{(1\times m)}{\mathbf{x}_{ikt}^{-i\prime}}, \underset{(1\times m)}{\mathbf{x}_{ikt-1}'}, \underset{(1\times m)}{\mathbf{z}_{ikt}'}, \underset{(1\times 1)}{1}\right)'$$

is the complete list of all exogenous variables, or IV,

$$\underset{((3m+3)\times 1)}{\mathbf{x}_{ik}} = \left(\underset{(1\times 1)}{y_{ik}^{-i}}, \underset{(1\times 1)}{y_{ik}^{t-1}}, \underset{(1\times m)}{\mathbf{x}_{ikt}'}, \underset{(1\times m)}{\mathbf{x}_{ikt}^{-i\prime}}, \underset{(1\times m)}{\mathbf{x}_{ikt-1}'}, \underset{(1\times 1)}{1}\right)'$$

is the complete list of all explanatory variables in the structural equation, and

$$\boldsymbol{\alpha} = \left(\alpha_1, \alpha_2, \boldsymbol{\alpha}_3', \boldsymbol{\alpha}_4', \boldsymbol{\alpha}_5', \alpha_6\right)$$

is the vector of coefficients.

The orthogonality between the IV and the structural error $\varepsilon_{ik} = \varepsilon_{ikt}$ implies the moment restriction

$$\mathbb{E}\left[\mathbf{z}_{ik}\varepsilon_{it}\right] = \mathbb{E}\left[\mathbf{g}_{ik}\left(\boldsymbol{\alpha}^*\right)\right] = \mathbf{0}_{(4m+2)\times 1}, \tag{5}$$

where $\boldsymbol{\alpha}^*$ is the "true" parameter in the structural equation. Estimation and statistical inference fall naturally within the GMM framework.

2.2 Testing Endogenous Effect under Weak Instruments

As discussed in the introduction, empirical evidence shows that the instruments in our data are weak for identifying α_1 in (4), creating bias and inconsistency in traditional IV estimation approaches. To better approximate the finite-sample behavior of two-stage least squares (2SLS) under weak IV, econometricians formulate the weak IV via a drifting sequence of models with the correlation coefficient between the endogenous variable and the IVs shrinking to zero at the rate n^{-b} (Andrews and Cheng, 2012, p.2165, Table 1). The strong IV corresponds to $b = 0$, under which the 2SLS estimator will converge in probability at the familiar rate $n^{-1/2}$. If the IV is "semi-strong", i.e. $0 < b < 1/2$, it will still be consistent but the convergence rate is slower than $n^{-1/2}$. When $b = 1/2$, it will be a non-degenerate random variable in the limit; in other words, it does not converge in probability to a constant no matter how large is the sample size. Lastly, if $b > 1/2$, 2SLS will be asymptotically unbounded in probability. The knife-edge case $b = 1/2$, the rate specified by Staiger and Stock (1997), is of particular theoretical interest, as it is the very rate that deprives 2SLS of estimation consistency but maintains its boundedness in probability. In a finite sample, weak instrumentation means that the point estimate via 2SLS could be misleading, for it may significantly deviate from the true value even if we have a large sample size.

Even though the presence of weak IV makes it difficult to estimate α_1 in (4), weak-IV robust tests are readily available to use in conducting inference under null hypotheses for the parameter of interest. In this paper, we employ two weak-IV robust tests applicable in the GMM framework, which allows dependence in clusters. We briefly summarize the logic of weak-IV robust tests in our context. Let

$$\underset{((3m+2)\times 1)}{\mathbf{z}_{ik}^{(1)}} = \left(\underset{(1\times 1)}{y_{ik}^{t-1}}, \underset{(1\times m)}{\mathbf{x}'_{ikt}}, \underset{(1\times m)}{\mathbf{x}_{ikt}^{-i\prime}}, \underset{(1\times m)}{\mathbf{x}'_{ikt-1}}, \underset{(1\times 1)}{1} \right)'$$

be the list of all included instruments, and $\boldsymbol{\eta} = (\alpha_2, \boldsymbol{\alpha}'_3, \boldsymbol{\alpha}'_4, \boldsymbol{\alpha}'_5, \alpha_6)$. Were α_1 known, the following equation

$$\mathbb{E}\left[\mathbf{z}_{ik}\left(y_{ik} - \alpha_1 y_k^{-i}\right)\right] = \mathbb{E}\left[\mathbf{z}_{ik}\left(\mathbf{z}_{ik}^{(1)\prime}\boldsymbol{\eta} + \varepsilon_{ik}\right)\right] = \mathbb{E}\left[\mathbf{z}_{ik}\mathbf{z}_{ik}^{(1)\prime}\right]\boldsymbol{\eta}, \qquad (6)$$

would over-identify $\boldsymbol{\eta}$. This is because $\mathbf{z}_{ik}^{(1)}$ is a proper subset of $\mathbf{z}_{ik}$ so that the rank condition is satisfied, and the additional variables from period $t-1$ provide extra moment restrictions. Therefore, under the null hypothesis

$$\mathrm{H}_0 : \alpha_1 = \alpha_1^*,$$

the standard GMM theory applies: $\boldsymbol{\eta}$ can be consistently estimated at the standard rate of convergence, the J-statistic follows a χ^2 asymptotic distribution, and the (scaled) sample average of the score function is asymptotically normal.

The well known Anderson and Rubin (1949) test (AR test) is based on value of a quadratic criterion function. Stock and Wright (2000) generalize the AR test to the Anderson-Rubin-type S-test (S-test, henceforth) under the GMM framework. The S-statistic is the value of the GMM criterion function from the continuous updating estimator (Hansen, Heaton, and Yaron, 1996). Under the null hypothesis, it follows a χ^2 distribution asymptotically, with the degrees of freedom being the number of over-identification constraints $\dim(\mathbf{z}_{ik}) - \dim\left(\mathbf{z}_{ik}^{(1)}\right) = (4m+2) - (3m+2) = m$. Alternatively, Kleibergen (2005)'s Lagrangian multiplier test (KLM test, henceforth) is based on the asymptotic distribution of the Lagrangian multiplier associated with the restriction $\alpha_1 = \alpha_1^*$, while the other coefficients are projected out as nuisance parameters. In our application, under the null the KLM test statistic follows $\chi^2(1)$ asymptotically.

The above reasoning justifies the size of the weak-IV robust tests. Regarding the power of the tests, if the IV turns out to be strong, the behavior of the test statistics can be derived by standard asymptotic analysis. Other than this favorable case, we again have to refer to the drifting sequence of models with the correlation coefficient shrinking at the rate n^{-b} (Andrews and Cheng, 2012). If $0 < b < 1/2$, the weak-IV robust tests are still consistent at any fixed alternative $\alpha_1 \neq \alpha_1^*$, meaning that the tests can reject the null under any fixed alternative with probability arbitrarily close to 1 if the sample size is sufficiently large. If $b \geq 1/2$, the tests still have power but are inconsistent; in other words, no matter how large is that sample size, there is non-zero probability of wrongly accepting the null when the null is indeed false. In the extreme case that the correlation coefficient is 0, i.e. the IVs are completely uninformative, the weak-IV tests have no power to detect any alternative.

So far we have laid out the structural model of endogenous effects, and the main idea of testing the effects under weak IV. Next, we introduce the background of our empirical study and summarize the data.

3 Data

China's Ministry of Education carried out the *Action Plan of Revitalizing Education in the 21st Century* in 1999, targeting at a significant boost of tertiary enrollment rate. To accommodate the swarm of new students, universities were encourage to setup multiple campuses. We examine a dataset of students in the School of Economics at an elite Chinese university. In 2004, his data-source university, located in the downtown of a provincial capital, opened a new campus in the city's outskirt. Our sample is based on the incoming freshman classes in 2006, 2007 and 2008. We observe administrative data of all the students in these cohorts for their freshman and sophomore year if the student was enrolled in the School of Economics for each of those years.

3.1 Room Assignment and Reshuffle

The university operates dormitories specifically for School of Economics students. Students are not allowed to make roommate requests in either their freshman or sophomore year, and assignment to rooms appears to be entirely ad-hoc and so potentially quasi-random. Each student takes a series of exams at the end of each semester. These exams are very high stakes: they determine students' grades and are critical for both merit-based scholarships and honors designation. The outcome variable (average test score) that we analyze is the average on a percentage scale of the final grades in all courses during the year. We observe this average test score for all students in our sample, though our data do not allow us to track students outside of the School of Economics.

School of Economics students in all three cohorts completed their freshmen year on the new campus and transfered to the old campus for their sophomore year. The dormitories on the old campus have primarily 4 person rooms, and the dormitories in the new campus have primarily 6 person rooms. However, the number of roommates varies across the sample even within the freshman or sophomore year. Not all rooms host 4 or 6 students.

The room information is compiled by the administration of the School of Economics. Our raw data of 2006, 2007, and 2008 freshmen in the School of Economics contain 797 students identified by a university-provided unique ID. Some students started in the School of Economics but are not observed in both years, because they either transferred into or out of the School, attended an overseas exchange program, or did not live in dormitories for at least one year. We also exclude those students who later transferred into the School of Economics, because those people typically have no room record in the first year, and they did not reside with students from School of Economics when they studied in other schools. These criteria restrict our sample to 757 students who were present in the sample for both years and have valid room numbers for both years. Freshman and sophomore year test scores and all covariates are observed for those 757 students in our sample.

As is discussed in Section 2.1, we further restrict our analysis to a subsample of treated individuals to implement our identification strategy. To construct the variable $\mathbf{z}_{ikt}$, among the 757 students we keep every individual who has at least one treatment individual and one instrumental individual. We identify a sample of 498 such treated individuals. The substantial loss of observations comes from rooms not split in the reassignment process. For example, when a year 1 four-people room is completely absorbed into a year 2 six-people room, the two new students in the year 2 six-people room have no instrumental individuals. They have to be excluded from the subsample of treated individuals.

[Table 1 about here.]

A second limitation of our data is the incomplete list of roommates. Although the freshman and sophomore dormitories have primarily 4 and 6 person rooms, in our data many freshmen do not have 3 roommates identified and, similarly, many sophomores do not have 5 roommates identified. The distribution of rooms for our population of 797 students is shown in Table 1—panel I for the freshman year and panel II for the sophomore year. In the freshman year, 744 students are in rooms with at least 3 people in our sample, but 53 of them have only 1 or 2 students in a room. Similarly, in the sophomore year, 674 students are in rooms with at least 4 residents identified in the data, but again a non-trivial fraction of students are in rooms with 3 or less people. This shortfall cannot be entirely explained by students whose room number is missing. The singleton observations suggest that either some individuals have errors in their room numbers in the administrative file or the dormitories had some rooms set aside as singles. Therefore, we further restrict our sample to treated individuals who we observe in relatively full rooms as sophomores, 4 or more total people, and instrumental individuals who we observe in relatively full rooms as freshman, 3 or more total people. This restriction leads to a final analysis sample of 464 students, which we call the *full room sample.* The observation loss is mild, as many students in small rooms are already dropped when collecting the treated individual subsample.

We impose this restriction for two reasons. First, we would like to compare the impact of endogenous peer effects in a relatively homogeneous settings, and this restriction eliminates individuals who are potentially in an unusual housing situation within their dormitory. Second, if the individuals are listed in a small population room due to a room number miscoding, then we avoid substantial measurement error in the identity of their peers arising from not observing most of their peers.

3.2 Summary Statistics

[Table 2 about here.]

Table 2 shows the means and standard deviations of our key variables for the entire freshman sample (797 observations), the sample of treated individuals (498 observations), and the full room sample (464 observations). At the treated student level, we observe their freshman and sophomore year test scores from 0 to 100,[5] their gender (binary, 1 for female and 0 for male), their ethnic group (1 for the Han Chinese, the dominant majority, and 0 for other minority ethnic groups), student-reported family income in 1,000's of Yuan, whether the student comes from a single parent family (binary), and whether they resided in the home province of the university or not prior to attending the school (binary).[6] We also observe the average current test scores and attributes of their freshman and sophomore year roommates, and the average attributes of the instrumental individuals for the sophomore year roommates' test scores.[7] Looking across the columns, the sample restrictions from the full freshman sample to the treated individuals and to the full room sample of treated individuals have virtually no effects on the means of our student test scores or attributes.

4 Empirical Analysis

Throughout this paper, in addition to the explicitly mentioned regressors we also control two fixed effect terms in all regression analyses: the dummy of *2006 Freshman Cohort* and the dummy of *2007 Freshman Cohort*. All standard errors and the variance components in test statistics are clustered at the dormitory room level, so that the inference is robust to within-room correlation of the unobservable disturbance.

4.1 Evidence on Random Assignment

We conduct a simple reduced-form falsification test to check the maintained presumption of random room assignment. Under random assignment, at year 1 a treated individual i has

[5] The university uses a percentage scale for test scores, and our average test score is the average across all percentage scores for the year.

[6] This last variable *Home Province* is important as the university enrolls about half of the students from the home province. Chinese university admission quotas are established province by province, and so two students from different provinces do not directly compete with each other.

[7] All rooms are single-sex, so gender does not appear in these averages attributes.

no connection with his instrumental individual, who only indirectly influences i's test score in year 2. As a result, the treated individual's year 1 test scores should be independent of any observable year 1 information concerning the instrumental individuals. We continue Example 1 to illustrate the implication.

Example 1 (continue)**.** (**Falsification test**) The period 1 outcome of a treated individual should be explained by the individual's own attributes and the attributes of his period 1 roommates. To falsify the claim of random assignment, we check the statistical association between the period 1 outcome and the attributes of the instrumental individuals. That is to say, under random assignment the coefficient $\boldsymbol{\gamma}_3$ in the regression equation

$$y_{21} = \boldsymbol{\gamma}_1'\mathbf{x}_2 + \boldsymbol{\gamma}_2'\frac{\mathbf{x}_6 + \mathbf{x}_7}{2} + \boldsymbol{\gamma}_3'\frac{\mathbf{x}_4 + \mathbf{x}_5}{2} + \gamma_4 + u_{21} \tag{7}$$

should be zeros. □

[Table 3 about here.]

Table 3 presents the regression and the falsification test for our full room sample of 464 observations.[8] We conduct a Wald test for the null that the coefficients on the 4 instruments are all zeros, that is, $\boldsymbol{\gamma}_3 = \mathbf{0}_{4\times1}$. The resulting Wald statistic is 4.847 and the p-value is 0.303. None of the coefficients on instrumental individual attributes are statistically significant at the 5% level.

[8] Guryan, Kroft, and Notowidigdo (2009) point out the bias in regressions caused by using the peer average that omits the subject individual as an explanatory variable. They argue intuitively that the bias is severe when an "urn", or a cluster, is small, and the bias is mitigated when the cluster grows bigger. However, Caeyers and Fafchamps (2016) formally investigate the source and magnitude of the bias. First, with a fixed sample size, they show analytically that the bias increases with the size of the peer group, presumably because a larger peer group implies a smaller number of groups. Further, while fixed effects exacerbate this bias, Caeyers and Fafchamps (2016) show that the bias vanishes as the fixed effect cluster size increases to infinity. Our data is consistent with minimal bias in that we have a substantial number of rooms in the sample thanks to the modest room sizes, and our fixed effects are captured at large clusters of the admission cohorts and the gender level. Caeyers and Fafchamps (2016)'s simulations suggest negligible bias given our sample, peer group and cluster sizes. Thus, we ignore the bias identified by Guryan, Kroft, and Notowidigdo (2009) throughout this paper.

Second, again in the full room sample, we conduct a standard balancing test. If the attributes of the instrumental individuals can explain the attributes of the treated individuals, it would suggest sorting into dormitory rooms as opposed to random assignment. We continue Example 1 to illustrate the balance test.

Example 1 (continue)**.** (**Balance test**) Since each individual has 4 attributes, namely *Han Chinese*, *Family Income*, *Single Parent Family* and *Home Province*, we write the regression in a four-equation simultaneous equations system

$$\underset{(4\times 1)}{\mathbf{x}_2} = \mathbf{A}_2 \cdot \frac{\mathbf{x}_6 + \mathbf{x}_7}{2} + \mathbf{A}_3 \cdot \frac{\mathbf{x}_4 + \mathbf{x}_5}{2} + \mathbf{a}_4 + \mathbf{u}_2^x,$$

where $\mathbf{A}_2$ and $\mathbf{A}_3$ are two 4×4 coefficient matrices, and $\mathbf{a}_4$ is a 4×1 coefficient vector. □

It is convenient to implement the standard Lagrangian multiplier (LM) test for the joint null hypothesis $\mathbf{A}_1 = \mathbf{A}_2 = \mathbf{0}_{4\times 4}$. With the variance clustered at the room level, the LM statistic is 40.305 with an associated p-value of 0.149. The test does not reject the null at the 10% level. This balance test gives additional supporting evidence of random room assignment.

4.2 Indirect Regression

The encouraging evidence of quasi-random assignment from the falsification test and the balancing test supports using our data to test for the existence of endogenous effects. In the context of Example 1, the existence of the endogenous peer effects can be examined by running an *indirect regression* or reduced form model: we regress y_{22} on the complete list of instruments

$$y_{22} = \delta_1 y_{21} + \boldsymbol{\delta}_2' \mathbf{x}_2 + \boldsymbol{\delta}_3' \frac{\mathbf{x}_1 + \mathbf{x}_3}{2} + \boldsymbol{\delta}_4' \frac{\mathbf{x}_6 + \mathbf{x}_7}{2} + \boldsymbol{\delta}_5' \frac{\mathbf{x}_4 + \mathbf{x}_5}{2} + \delta_6 + \varepsilon_{22}. \tag{8}$$

Notice that $(\mathbf{x}_4 + \mathbf{x}_5)/2$ is the average attributes of person 2's instrumental individuals 4 and 5. As is clear from (3), the instruments enter the expression for y_{22} in (8) only if α_1 in (1) is non-zero. Thus, the value of $\boldsymbol{\delta}_5$ provides indirect evidence about α_1. If $\alpha_1 = 0$, then $\boldsymbol{\delta}_5 = \mathbf{0}_{4\times1}$; conversely, $\boldsymbol{\delta}_5 \neq \mathbf{0}_{4\times1}$ implies $\alpha_1 \neq 0$.

[Table 4 about here.]

The first column of Table 4 presents the result of the indirect regression as in (8). The coefficients on *Home Province* and *Han Chinese* are both significant at the 5% level. Moreover, we reject the joint hypothesis $\boldsymbol{\theta}_5 = \mathbf{0}_{4\times1}$ at 5% level based on the Wald test statistic 10.628. This regression alludes to the existence of endogenous effects; however, the test of the null for $\boldsymbol{\theta}_5$ does not directly evaluate α_1. We would prefer direct estimation and inference concerning the endogenous effects if possible.

4.3 Evidence of Weak Instruments

2SLS is the most popular procedure to estimate a linear structural model with excluded IVs. In view of the potential within-room dependence, we employ the two-step GMM to achieve asymptotic efficiency. The second column of Table 4 displays the two-step GMM estimates based on the moment restriction (5). An efficient GMM weighting matrix is constructed by using the 2SLS estimate as the preliminary estimator.

To our surprise, the instrumental variable estimate of the effect of sophomore roommate's test score is quite large at 0.713. This means keeping everything else the same, a student's test score would enjoy a 0.713 percentage point hike if his roommates' average test score increases by 1 point. The point estimate even exceeds the estimated direct effect of the individual's own freshmen test score 0.651, which presumably acts as an effective proxy for individual ability. Such a large endogenous peer effect seems unlikely.

Our first reaction was: did we misspecified the model? Perhaps the moment conditions simply fail, so the "instruments" are correlated with the structural error in (1). The *J*-

statistic of 3.965 and the corresponding p-value of 0.265, however, suggests that it passes the over-identification test. There is no statistical evidence of violation of the moment restrictions.

Besides orthogonality, which the J-statistics has just checked, the other pillar for the standard asymptotic theory of IV methods is the so-called relevance condition. The relevance calls for non-trivial correlation between the endogenous variable and the instruments. We examine the first-stage regression of the sophomore year average roommate test score on all the instruments. As is reported in the third column of Table 4, the message is mixed. On the one hand, three coefficients out of the four excluded instruments are statistically significant at the 5% level. For the joint test of the null “the coefficients of all four excluded instruments are zeros”, the Wald statistic is 18.024, which strongly rejects the null. This is encouraging, indicating that the excluded instruments have explanatory power, and is consistent with our presumption that peer effects from the freshman year roommates influence the second year test scores. On the other hand, statistical significance is not sufficient to avoid the weak IV problem in estimation. In fact, the F-statistic associated with this set of excluded instruments is only 6.337, which is well below the widely used threshold that assures the standard asymptotic theory of IV estimation (Staiger and Stock, 1997).[9]

The first-stage results raise concerns over weak instruments as a plausible explanation for the very high estimated effect of the sophomore roommate's test score. We now turn to our properly sized tests developed by Stock and Wright (2000) and Kleibergen (2005) to evaluate the endogenous peer effects. The idea of weak-IV robust tests has been introduced in Section 2.2. In this following two subsections, we describe the procedures formally.

4.4 Anderson-Rubin-Type S-Test

We implement the continuous updating estimator (CUE), an information-theoretical variant of GMM, with a given α_1. Let $n = \sum_{k=1}^{K} r_{kt}$ be the total number of qualified individuals,

[9] See Olea and Pflueger (2013) for the latest work of a test robust to cluster dependence.

and $\underset{(n\times(3m+2))}{\mathbf{Z}^{(1)}}$ and $\underset{(n\times(4m+2))}{\mathbf{Z}}$ be the data matrices of $\mathbf{z}_{ik}^{(1)}$ and $\mathbf{z}_{ik}$, respectively. Let

$$\bar{\mathbf{g}}_n(\boldsymbol{\alpha}) = n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{r_{kt}}\mathbf{g}_{ik}(\boldsymbol{\alpha})$$

be the sample analogy of the moment condition evaluated at $\boldsymbol{\alpha}$. To explicitly take into account the clustering of the structural error, we estimate the asymptotic variance of $\sqrt{n}\bar{\mathbf{g}}_n(\boldsymbol{\alpha})$ as

$$\widehat{\boldsymbol{\Omega}}_n(\boldsymbol{\alpha}) = n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{r_{kt}}\sum_{j=1}^{r_{kt}}\omega_{ijk}\left(\mathbf{g}_{ik}(\boldsymbol{\alpha}) - \bar{\mathbf{g}}_n(\boldsymbol{\alpha})\right)\left(\mathbf{g}_{jk}(\boldsymbol{\alpha}) - \bar{\mathbf{g}}_n(\boldsymbol{\alpha})\right)',$$

where the kernel function $\omega_{ijk} = \mathbf{1}\{i=j\} + (r_{kt}-1)^{-1}\mathbf{1}\{i\neq j\}$. We set up the GMM criterion function

$$J(\boldsymbol{\alpha}) = n\bar{\mathbf{g}}_n(\boldsymbol{\alpha})'\widehat{\mathbf{W}}_n(\boldsymbol{\alpha})\bar{\mathbf{g}}_n(\boldsymbol{\alpha}),$$

and $\widehat{\mathbf{W}}_n(\boldsymbol{\alpha}) = \widehat{\boldsymbol{\Omega}}_n^{-1}(\boldsymbol{\alpha})$. Under a given α_1, CUE estimates

$$\tilde{\boldsymbol{\eta}}(\alpha_1) = \arg\min_{\eta} J\left((\alpha_1, \boldsymbol{\eta}')'\right).$$

The logic behind the Anderson-Rubin-type S-test is similar to the original AR test. Given an α_1, we can rewrite the structural equation for individual i in room k as

$$y_{ik} - \alpha_1 y_{ik}^{-i} = \boldsymbol{\eta}'\mathbf{z}_{ik}^{(1)} + \varepsilon_{ik},$$

in which no endogenous variables appear on the right-hand side of the above equation. The asymptotic theory of the GMM estimator for $\boldsymbol{\eta}$ is standard. Since the complete list of IV $\mathbf{z}_{ik}$ includes all variables in $\mathbf{z}_{ik}^{(1)}$, the coefficient $\boldsymbol{\eta}$ is consistently estimable. The excluded instruments, in addition to improving estimation efficiency, provide over-identification restrictions to test IV orthogonality. The S-statistic (Stock and Wright, 2000) is nothing but the J-statistic of CUE evaluated at $\tilde{\boldsymbol{\alpha}} = \left(\alpha_1, \tilde{\boldsymbol{\eta}}\left(\alpha_1\right)'\right)'$, that is,

$$\mathcal{S}\left(\alpha_1\right) = J\left(\tilde{\boldsymbol{\alpha}}\right).$$

If $\alpha_1 = \alpha_1^*$, asymptotically

$$\mathcal{S}\left(\alpha_1^*\right) \xrightarrow{d} \chi^2\left(m\right).$$

where the degrees of freedom of the χ^2 distribution equals the number of excluded instruments. In our empirical context $m = 4$. This asymptotic distribution holds no matter how weak is the instrumentation.

4.5 Kleibergen's Lagrangian Multiplier Test

In addition to the Anderson-Rubin-type S-test, the KLM test also provides an alternative test under weak IV. The KLM test is based on the behavior of the score of $J\left(\boldsymbol{\alpha}\right)$:

$$\mathbf{q}_n\left(\boldsymbol{\alpha}\right) = n^{-1/2}\frac{\partial J\left(\boldsymbol{\alpha}\right)}{\partial \boldsymbol{\alpha}} = -2\widehat{\boldsymbol{\Sigma}}_n'\widehat{\mathbf{W}}_n\left(\boldsymbol{\alpha}\right)\sqrt{n}\bar{\mathbf{g}}_n\left(\boldsymbol{\alpha}\right) + o_{\mathrm{p}}\left(1\right).$$

where $\underset{((4m+2)\times(3m+3))}{\widehat{\boldsymbol{\Sigma}}_n} = n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{r_{kt}}\mathbf{z}_{ik}\mathbf{x}_{ik}'$ due to the linearity of $\bar{\mathbf{g}}_n\left(\boldsymbol{\alpha}\right)$ in our context. In the strong identification case, $\widehat{\boldsymbol{\Sigma}}_n$ converges in probability to a fixed full-rank non-random matrix. Under the null hypothesis $\alpha_1 = \alpha_1^*$, the score $\mathbf{q}_n\left(\tilde{\boldsymbol{\alpha}}^*\right)$ is asymptotically normal. However, when $\widehat{\boldsymbol{\Sigma}}_n$ is asymptotically rank deficient, the asymptotic distribution of $\mathbf{q}_n\left(\tilde{\boldsymbol{\alpha}}^*\right)$ is generally unknown due to the dependence between $\sqrt{n}\bar{\mathbf{g}}_n\left(\tilde{\boldsymbol{\alpha}}^*\right)$ and $\widehat{\boldsymbol{\Sigma}}_n$.

To derive a pivotal statistic robust to the potential rank deficiency of $\widehat{\boldsymbol{\Sigma}}_n$, Kleibergen (2005) proposes an alternative estimator $\underset{(4m+2)\times(3m+3)}{\widehat{\mathbf{D}}\left(\boldsymbol{\alpha}\right)}$ for the Jacobian $\mathbb{E}\left[\partial\bar{\mathbf{g}}_n\left(\boldsymbol{\alpha}\right)/\partial\boldsymbol{\alpha}'\right]$. In our linear model, the vectorized $\widehat{\mathbf{D}}\left(\boldsymbol{\alpha}\right)$ can be computed by

$$\operatorname{vec}\left(\widehat{\mathbf{D}}\left(\boldsymbol{\alpha}\right)\right) = n^{-1}\mathbf{V}_n'\mathbf{M}_{\mathbf{G}(\boldsymbol{\alpha})}\mathbf{1}_n,$$

where $\text{vec}(\cdot)$ is the column-wise matrix vectorization, $\underset{(n\times((4m+2)(3m+3)))}{\mathbf{V}_n}$ is the matrix whose each row equal to $(\mathbf{x}_{ik}\otimes\mathbf{z}_{ik})'$, and

$$\mathbf{M}_{\mathbf{G}(\boldsymbol{\alpha})}=\mathbf{I}_n-\mathbf{G}(\boldsymbol{\alpha})\left(\mathbf{G}(\boldsymbol{\alpha})'\mathbf{G}(\boldsymbol{\alpha})\right)^{-1}\mathbf{G}'(\boldsymbol{\alpha}),$$

where $\underset{n\times(4m+2)}{\mathbf{G}(\boldsymbol{\alpha})}$ is the matrix whose each row equal to $\mathbf{g}_{ik}(\boldsymbol{\alpha})'$. The matrix $\widehat{\mathbf{D}}(\boldsymbol{\alpha})$ is particularly constructed to project out the influence of $\bar{\mathbf{g}}_n(\boldsymbol{\alpha})$ in order to achieve asymptotically independence from $\sqrt{n}\bar{\mathbf{g}}_n(\boldsymbol{\alpha})$. Given these components, the KLM statistic is defined as

$$\mathcal{K}(\alpha_1)=n\bar{\mathbf{g}}_n(\tilde{\boldsymbol{\alpha}})'\widehat{\mathbf{W}}_n(\tilde{\boldsymbol{\alpha}})\widehat{\mathbf{D}}(\tilde{\boldsymbol{\alpha}})\left(\widehat{\mathbf{D}}(\tilde{\boldsymbol{\alpha}})'\widehat{\mathbf{W}}_n(\tilde{\boldsymbol{\alpha}})\widehat{\mathbf{D}}(\tilde{\boldsymbol{\alpha}})\right)^{-1}\widehat{\mathbf{D}}(\tilde{\boldsymbol{\alpha}})'\widehat{\mathbf{W}}_n(\tilde{\boldsymbol{\alpha}})\bar{\mathbf{g}}_n(\tilde{\boldsymbol{\alpha}}).$$

If α_1 is strongly identified, under standard assumptions the conventional LM test statistic converges in distribution to the χ^2 distribution with the degrees of freedom being the number of restrictions, i.e. $\chi^2(1)$ in our context. The difficulty in the weak IV scenario is the asymptotic rank deficiency of $\widehat{\boldsymbol{\Sigma}}_n$. Using $\widehat{\mathbf{D}}(\boldsymbol{\alpha})$ to replace $\widehat{\boldsymbol{\Sigma}}_n$, Kleibergen (2005) shows that if $\alpha_1=\alpha_1^*$, we have

$$\mathcal{K}(\alpha_1^*)\xrightarrow{d}\chi^2(1).$$

Next, we will apply both the S-test and the KLM test to our dataset. A careful reader might consider the Wald test as a potential third alternative test. However, the asymptotic distribution of the Wald statistic, which is based on an unrestricted estimator, depends on the strength of the correlation between the endogenous variable and the IV. When instruments are weak, the asymptotic distribution of the Wald statistic is non-pivotal, rather than being distributed $\chi^2(1)$ (Dufour, 1997; Staiger and Stock, 1997; Zivot, Startz, and Nelson, 1998; Moreira, 2003). Another popular weak-IV robust test is Moreira (2003)'s conditional likelihood ratio statistic (CLR) under the independently and identically distributed setting. Kleibergen (2005, p.1113) adapts CLR to the *conditional GMM statistic*. However, due to the non-standard asymptotic distribution of the GMM version of CLR, which has to be numerically simulated, we will not implement the CLR test in this paper.

4.6 Results

We conduct the weak-IV robust tests for a sequence of hypothesized values of the endogenous test score effects ranging from -0.1 to 2.0. Since the two tests follow different asymptotic null distributions, we do not directly compare the test statistics. Instead, Figure 2 shows the p-values of the S-test and the KLM test at different values of the hypothesized endogenous effect α_1. The horizontal dashed lines are 0.01, 0.05 and 0.10, the commonly used test sizes. In the upper graph α_1 ranges from -0.1 to 2, and the lower graph zooms in the shaded region where $\alpha_1 \in [-0.1, 0.3]$. For any null value, the reader can identify that value on the horizontal axis, and then look upwards to the solid or dashed lines in order to read the p-value of the vertical axis and so determine whether this particular null can be rejected with confidence.

The null hypothesis of zero or no endogenous effects is rejected at 5% size for both the the S-test and the KLM test. At $\alpha_1 = 0$, the S-test and the KLM test statistics are 11.523 and 5.381, respectively; the p-value curves, accordingly, cross the vertical line of $\alpha_1 = 0$ at p-value 0.0213 and 0.0203, respectively.

Next, we can invert the S-test or the KLM test to construct confidence intervals of desirable asymptotic coverage probability. We can therefore use this figure to back out the lower bounds of such confidence intervals. For example, Figure 2 shows that the 90% confidence interval of the KLM test is $[0.28, 1.51]$, and the 95% confidence interval is $[0.15, 1.68]$. Similar confidence intervals are given by inverting the S-test, as its p-value curve also crosses the 0.05 horizontal line at around $\alpha = 0.16$. The evidence suggests positive endogenous test score peer effects are sizable in our data. In comparison, the IV 95% confidence lower bound constructed from information in Table 4 is 0.217, which is 45% larger than the weak-IV robust lower bound.[10] The upper bound of the confidence interval is less informative here;

[10] Table 4 reports the point estimate 0.715 and standard error 0.254 in the two-step GMM estimation of the

the larger-than-1 endogenous effects contradict economic common sense, and again indicate the problem of weak IV.

[Figure 2 about here.]

We are also interested in the effect of the other explanatory variables. Unlike the strongly identified case, here under the presence of weak IV, these coefficients cannot be consistently estimated. As a suboptimal solution, we consider the coefficient $\tilde{\boldsymbol{\eta}} = \tilde{\boldsymbol{\eta}}\left(\alpha_1\right)$ under various hypothesized values of α_1. From the statistical evidence and economic common sense, we restrict $\alpha_1 \in [0, 1]$, a reasonable range. If some coefficients significantly differ from zero under a wide range of feasible values of α_1, we have strong evidence about the effect of the corresponding variable. Figure 3 displays the point estimate of each element in $\tilde{\boldsymbol{\eta}}$ (the solid line) and the 95% confidence interval (the upper and lower dash lines). The confidence interval is constructed by the standard formula, that is, the point estimate plus and minus 1.96 times the standard error.

The graph suggests that the *Freshman Test Score*, *Home Province*, and the dummy of *2006 Freshman Cohort* are significantly different from 0. It is not surprising that the own freshman year test score is important in explaining the sophomore year test score. The significance of home province is also reasonable, because college students from that particular home province have earned a national-wide reputation for academic excellence—a blessing, as well as a curse, of ferocious competition in China's College Entrance Examination. In the meantimes, *Gender*, *Sophomore Roommate Family Income*, the dummy of *2007 Freshman Cohort*, and the intercept are significant only if the true α_1 is close to 0, say, between 0 and 0.15. All the freshman roommate attributes are individually insignificant, which is consistent with our observation of weak IV: the freshman roommate attributes are only weakly related to a person's sophomore test score.

[Figure 3 about here.]

coefficient of sophomore roommate test score. Accordingly, the lower bound of the 95% confidence interval is $0.715 - 1.96 \times 0.254 = 0.217$.

4.7 Robustness Check

In order to check the robustness of our result, we run another weak-IV robust test similar in spirit to the falsification test in Section 4.1. We replace $\mathbf{y}_{kt}$ by $\mathbf{y}_{kt-1}$ in (4) and rewrite the structural equation as

$$\mathbf{y}_{kt-1} = \phi_1 \mathbf{y}_{kt}^{-i} + \mathbf{X}_{kt}\mathbf{\Phi}_2 + \mathbf{X}_{kt}^{-i}\mathbf{\Phi}_3 + \mathbf{X}_{kt-1}^{-i}\mathbf{\Phi}_4 + \phi_5 \mathbf{I}_{r_{kt}} + \boldsymbol{\varepsilon}_{kt-1}.$$

We use the same excluded variables $\mathbf{z}_{ikt}$ to instrument the endogenous variable $\mathbf{y}_{kt}^{-i}$. If the room assignment is random, we expect that ϕ_1 to be zero because future roommates should not affect pre-determined test scores.

[Figure 4 about here.]

According to Figure 4, the p-value of the weak-IV tests shows no evidence to reject the null of zero endogenous effects. Specifically, the p-value curves reach a maximum near the null of $\phi_1 = 0$, and rejection of a hypothesized null only arises for endogenous effects that are well away from zero. This result supports the presumption of random room assignment from another aspect.

5 Summary and Conclusion

This paper exploits a unique dataset from a Chinese University that allows us to examine the endogenous effects of college roommates on academic performance. We utilize a recently developed strategy for identifying endogenous spillovers. In partially overlapping peer groups, the previous peers of one student can act as an instrument for that student's outcomes when examining the factors that determine the outcomes of the student's current peers. In our context, the partially overlapping peer groups arise because the university housed incoming freshman at a new campus prior to moving those students to the old campus for their sophomore year, while dormitory room sizes are different in the two campuses. Balancing and falsification tests support the maintained hypothesis of random assignment to new second year roommates: neither the attributes nor the first year test scores of students can be explained by the information on the first year roommates of those students' new second year roommates.

The major obstacle that we face in evaluating the endogenous effects between roommates is that exogenous information on the first year roommates are weak instruments for the test scores of second year students. Accordingly, we exploit weak-IV robust tests to test for the existence of endogenous effects. The size of these tests are correct under a null hypothesis concerning the magnitude of the endogenous effects. The probability of falsely rejecting the null of zero endogenous effects is about 0.02, while the lower bound of the 95% confidence interval is 0.15. Our estimates imply sizable endogenous effects, but is noticeably smaller than the lower bound of the naive confidence interval according to the conventional IV estimation. Finally, as another falsification test, we repeat our weak-IV robust tests for whether the second year roommates have endogenous effects on the student's first year test scores. The weak-IV robust approach cannot reject the null of zero endogenous effects on the pre-determined outcomes. These findings provide strong and relatively unique quasi-experimental evidence on the existence and size of endogenous peer effects on academic outcomes, and the techniques described above represent valuable tools for future studies that examine endogenous effects.

References

Aizer, A., and J. Currie (2004): "Networks or neighborhoods? Correlations in the use of publicly-funded maternity care in California," *Journal of Public Economics*, 88(12), 2573–2585.

Anderson, T. W., and H. Rubin (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations," *The Annals of Mathematical Statistics*, 20(1), 46–63.

Andrews, D., and X. Cheng (2012): "Estimation and inference with weak, semi-strong, and strong identification," *Econometrica*, 80(5), 2153–2211.

Angrist, J. D. (2014): "The perils of peer effects," *Labour Economics*, 30, 98–108.

Bayer, P., S. L. Ross, and G. Topa (2008): "Place of work and place of residence: informal hiring networks and labor market outcomes," *Journal of Political Economy*, 116(6), 1150–1196.

Bertrand, M., E. F. Luttmer, and S. Mullainathan (2000): "Network effects and welfare cultures," *The Quarterly Journal of Economics*, 115(3), 1019–1055.

Bifulco, R., J. M. Fletcher, and S. L. Ross (2011): "The effect of classmate characteristics on post-secondary outcomes: Evidence from the Add Health," *American Economic Journal: Economic Policy*, 3(1), 25–53.

Billings, S. B., D. J. Deming, and S. L. Ross (2016): "Partners in crime: schools, neighborhoods and the formation of criminal networks," Discussion paper, NBER Working Papers No.21962.

Brock, W., and S. Durlauf (2001): "Interactions-based models," *Handbook of Econometrics*, 5, 3297–3380.

Caeyers, B., and M. Fafchamps (2016): "Exclusion bias in the estimation of peer effects," Discussion paper, NBER Working Paper No.22565.

Carrell, S., R. L. Fullerton, and J. West (2009): "Does your cohort matter? Measuring peer effects in college achievement," *Journal of Labor Economics*, 27(3), 439–464.

De Giorgi, G., M. Pellizzari, and S. Redaelli (2009): "Be as careful of the company you keep as of the books you read: peer effects in education and on the labor market," Discussion paper, NBER Working Papers No.14948.

——— (2010): "Identification of social interactions through partially overlapping peer groups," *American Economic Journal: Applied Economics*, 2(2), 241–275.

De Giorgi, G., M. Pellizzari, and W. G. Woolston (2012): "Class size and class heterogeneity," *Journal of the European Economic Association*, 10(4), 795–830.

Dufour, J.-M. (1997): "Some impossibility theorems in econometrics with applications to structural and dynamic models," *Econometrica*, 65(6), 1365–1387.

Durlauf, S. N. (2004): "Neighborhood effects," in *Handbook of regional and urban economics*, ed. by J. Henderson, and J. Thisse, vol. 4, pp. 2173–2242. Elsevier.

Fletcher, J. M., and S. L. Ross (2012): "Estimating the effects of friendship networks on health behaviors of adolescents," Discussion paper, NBER Working Papers No.18253.

Foster, G. (2006): "It's not your peers, and it's not your friends: Some progress toward understanding the educational peer effect mechanism," *Journal of Public Economics*, 90(8-9), 1455–1475.

Friesen, J., and B. Krauth (2011): "Ethnic enclaves in the classroom," *Labour Economics*, 18(5), 656–663.

Glaeser, E. L., B. I. Sacerdote, and J. A. Scheinkman (2003): "The social multiplier," *Journal of the European Economic Association*, 1(2-3), 345–353.

Gould, E. D., V. Lavy, and M. Daniele Paserman (2009): "Does immigration affect the long-term educational outcomes of natives? Quasi-experimental evidence," *The Economic Journal*, 119(540), 1243–1269.

Guryan, J., K. Kroft, and M. J. Notowidigdo (2009): “Peer effects in the workplace: Evidence from random groupings in professional golf tournaments,” *American Economic Journal: Applied Economics*, 1(4), 34–68.

Hansen, L., J. Heaton, and A. Yaron (1996): “Finite-sample properties of some alternative GMM estimators,” *Journal of Business & Economic Statistics*, 14(3), 262–280.

Ioannides, Y. M., and L. Datcher Loury (2004): “Job information networks, neighborhood effects, and inequality,” *Journal of Economic Literature*, 42(4), 1056–1093.

Kleibergen, F. (2005): “Testing parameters in GMM without assuming that they are identified,” *Econometrica*, 73(4), 1103–1123.

Lavy, V., M. D. Paserman, and A. Schlosser (2012): “Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom,” *The Economic Journal*, 122(559), 208–237.

Lavy, V., and A. Schlosser (2011): “Mechanisms and impacts of gender peer effects at school,” *American Economic Journal: Applied Economics*, 3(2), 1–33.

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

Marmaros, D., and B. Sacerdote (2002): “Peer and social networks in job search,” *European Economic Review*, 46(4), 870–879.

——— (2006): “How do friendships form,” *The Quarterly Journal of Economics*, 121(1), 79–119.

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

Moreira, M. J. (2003): “A conditional likelihood ratio test for structural models,” *Econometrica*, 71(4), 1027–1048.

Olea, J. L. M., and C. Pflueger (2013): "A robust test for weak instruments," *Journal of Business & Economic Statistics*, 31(3), 358–369.

Ross, S. L. (2011): "Social interactions within cities: Neighborhood environments and peer relationships," in *Handbook Urban Economics Planning*, ed. by N. Brooks, K. Donaghy, and G. Knapp, pp. 203–229. Oxford University Press.

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

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557–586.

Stinebrickner, R., and T. R. Stinebrickner (2006): "What can be learned about peer effects using college roommates? Evidence from new survey data and students from disadvantaged backgrounds," *Journal of Public Economics*, 90(8), 1435–1454.

Stock, J. H., and J. H. Wright (2000): "GMM with weak identification," *Econometrica*, 68(5), 1055–1096.

Stock, J. H., J. H. Wright, and M. Yogo (2014): "A survey of weak instruments and weak identification in generalized method of moments," *Journal of Business & Economic Statistics*, 20(4), 518–529.

Stock, J. H., and M. Yogo (2005): "Testing for weak instruments in linear IV regression," in *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, ed. by J. Stock, and D. Andrews, pp. 80–108. Cambridge University Press.

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

Weinberg, B. A. (2007): "Social interactions with endogenous associations," Discussion paper, NBER Working Papers No.13038.

Zivot, E., R. Startz, and C. R. Nelson (1998): "Valid confidence intervals and inference in the presence of weak instruments," *International Economic Review*, 39(4), 1119–1144.

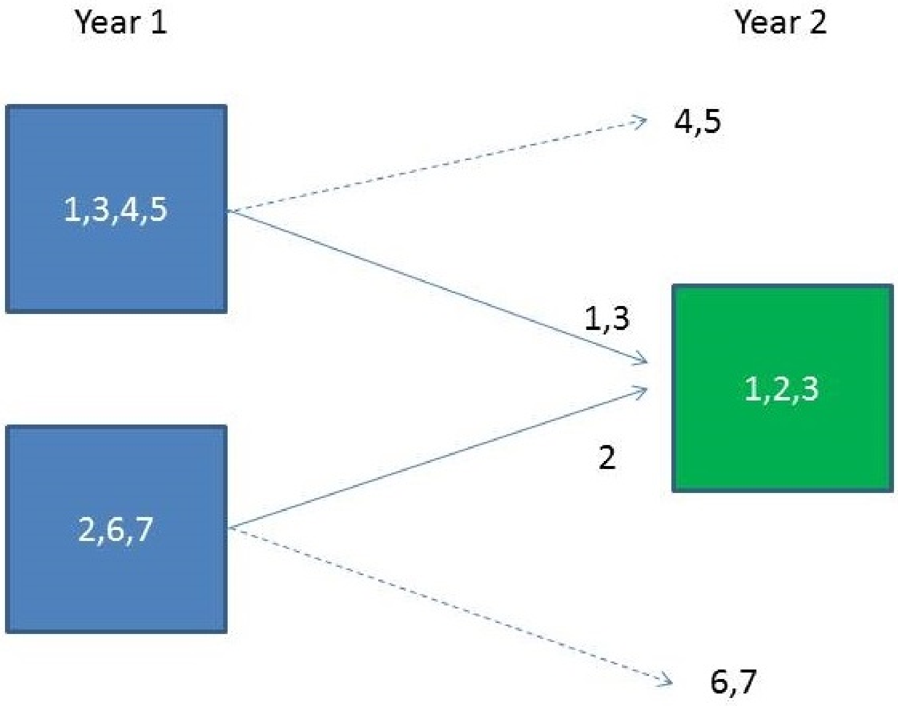
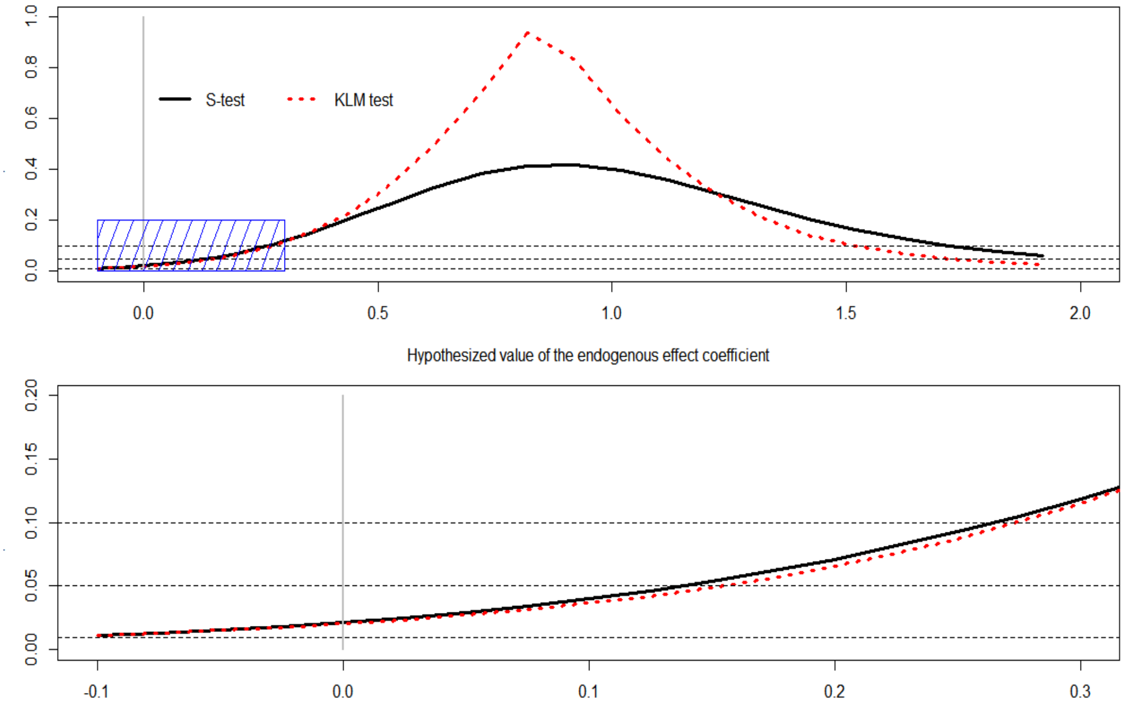
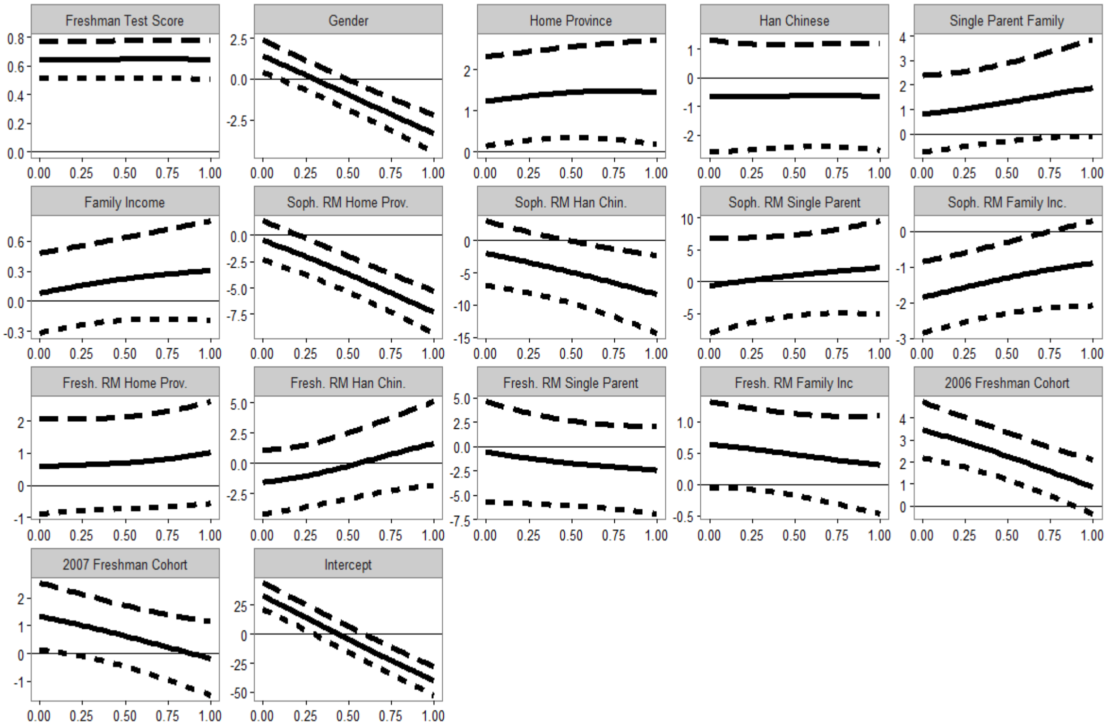


Figure 1: Room Reshuffle in Example 1



Notes: The two graphs display the p-value of the S-test (black solid curve) and the KLM test (red dotted curve). The X-axis shows the hypothesized value of α_1, and the Y-axis is the p-value. The upper graph shows the entire picture with α_1 ranging from -0.1 to 2, and the lower graph zooms in to show the region around $\alpha_1 = 0$, the shaded area in the upper graph.

Figure 2: p-values of the S-Test and the KLM Test



Notes: The X-axis displays the hypothesized value of α_1, and the Y-axis is the value of the coefficients in $\boldsymbol{\eta}$. The solid curve is the point estimate, and the dashed curves are the upper bound and the lower bound of the 95% confidence interval computed by the standard formula.

Figure 3: Confidence Intervals of Coefficients in $\boldsymbol{\eta}$ under Hypothesized Values of α_1

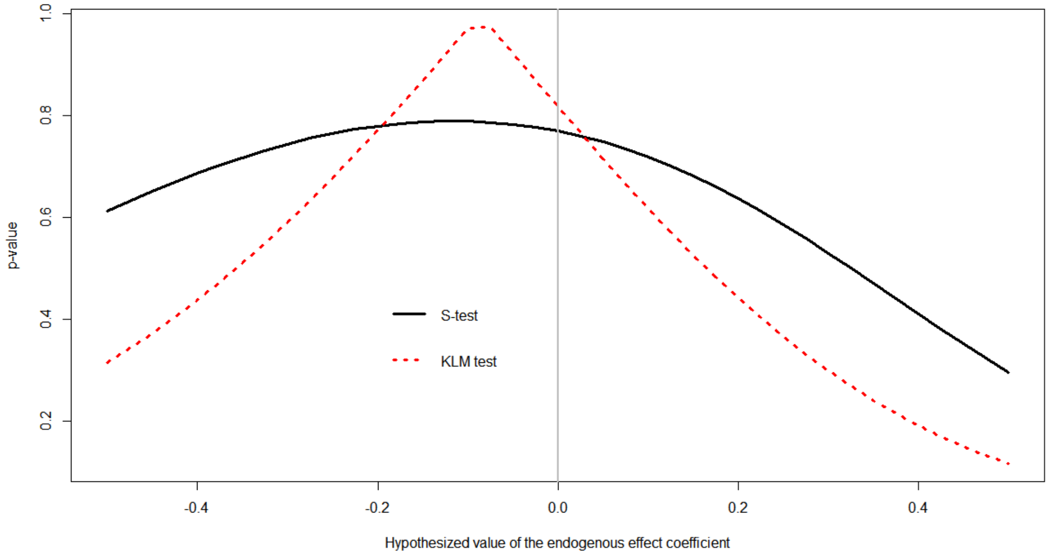


Figure 4: p-value of the S-Test and the KLM Test for Robustness Check

Table 1: Room Size Distributions

Panel I: Freshman Year Room Distributions

No. of residents in a room	missing	1	2	3	4	5	Total
Number of rooms	NA	35	18	69	124	1	262
Number of people	18	35	36	207	496	5	797
Subsample percent	11.16%			88.84%			100%

Panel II: Sophomore Year Room Distributions

No. of residents in a room	missing	1	2	3	4	5	6	7	Total
Number of rooms	NA	10	15	10	25	44	52	6	179
Number of people	53	10	30	30	100	220	312	42	797
Subsample percent	15.43%				84.56%				100%

Notes: Panels I and II are based on the full sample of all students observed across the three cohorts in their freshman and sophomore year, respectively. The first row represents the number of rooms with each number of residents at that time, and the second row calculates the number of students in rooms containing that number of residents. The subsample percentages are calculated for two subsamples in each panel—the total number of students in rooms that do not meet the criteria for the big room sample versus those in rooms that meet the criteria.

Table 2: Variable Means and Standard Deviations in Student Samples

Variable Names	Freshmen Sample		Treated Sample		Full Room Sample	
	mean	s.d.	mean	s.d.	mean	s.d.
Sophomore Test Score	82.071	7.988	81.834	8.397	82.010	8.348
Freshman Test Score	80.252	8.441	79.878	8.444	79.963	8.388
Gender	0.561	0.497	0.552	0.498	0.556	0.497
Han Chinese	0.910	0.287	0.906	0.293	0.905	0.293
Family Income	1.049	1.063	1.055	1.003	1.047	1.023
Single Parent Family	0.039	0.194	0.040	0.197	0.041	0.198
Home Province	0.494	0.500	0.480	0.500	0.487	0.500
2006 Freshman Cohort	0.338	0.473	0.325	0.469	0.349	0.477
2007 Freshman Cohort	0.299	0.458	0.283	0.451	0.269	0.444
Soph. Roommate Test	82.037	5.334	81.614	5.620	81.726	5.639
Soph. RM Han Chin.	0.912	0.149	0.913	0.143	0.911	0.145
Soph. RM Family Inc.	1.050	0.571	1.024	0.514	1.014	0.520
Soph. RM Single Parent	0.038	0.095	0.036	0.088	0.038	0.091
Soph. RM Home Prov.	0.518	0.292	0.506	0.288	0.500	0.283
Fresh. RM Han Chin.	0.910	0.193	0.909	0.179	0.910	0.170
Fresh. RM Family Inc.	1.044	0.727	1.092	0.733	1.077	0.733
Fresh. RM Single Parent	0.039	0.120	0.039	0.121	0.040	0.125
Fresh. RM Home Prov.	0.494	0.382	0.484	0.377	0.485	0.377
IV Han Chinese	NA	NA	0.885	0.235	0.884	0.237
IV Family Income	NA	NA	0.980	0.760	0.969	0.762
IV Single Parent Family	NA	NA	0.039	0.151	0.042	0.156
IV Home Province	NA	NA	0.499	0.372	0.497	0.374
Sample Size	797		498		464	

Notes: The columns contain the means and standard deviations for (in order from left to right) all freshmen present in the sample, all students present in both the freshman and sophomore samples with an instrumental individual (treated individuals), and finally the subsample of treated individuals who reside in relatively full rooms, i.e. three or more residents for freshman year and four or more residents for sophomore year. The variables listed from top to bottom begin with the attributes of the student, then the mean attributes of the student's sophomore and freshman roommates, and the average attributes of the freshman roommates of the sophomore roommates of the treated individuals (instrumental individuals).

Table 3: Falsification Test

Variable Names	estimate	s.e.
Gender	5.0853	0.7682
Han Chinese	7.6119	2.0100
Family Income	-0.3013	0.2829
Single Parent Family	0.0227	2.2789
Home Province	5.3910	0.7093
Intercept	62.6777	3.3646
2006 Freshman Cohort	-1.4504	0.9212
2007 Freshman Cohort	-0.7741	0.9319
Fresh. RM Han Chinese	3.1190	2.1828
Fresh. RM Family Inc.	0.5712	0.4841
Fresh. RM Single Parent	3.8617	2.6310
Fresh. RM Home Prov.	1.9982	0.9582
IV Han Chinese	0.2087	1.4265
IV Family Income	0.3006	0.4068
IV Single Parent Family	-0.2061	2.5708
IV Home Province	1.8632	1.0201
Wald Test for IV (p-value)	4.8472 (0.3033)	

Notes: the columns contain the estimates and standard errors for the regression of treated individual freshman year test scores on the treated individual attributes, freshman roommate attributes and instrumental individual attributes.

Table 4: Indirect Regression and Two-Stage Regression

	indirect reg.		Two-step GMM		1st-stage	
	estimate	s.e.	estimate	s.e.	estimate	s.e.
Soph. Roommate Test	NA	NA	0.7148	0.2545	NA	NA
Freshman Test Score	0.7171	0.0715	0.6503	0.0695	0.0595	0.0281
Gender	1.3148	0.5142	-1.9897	1.2688	4.5159	0.5971
Han Chinese	0.1542	1.0175	-0.5987	0.9330	0.6025	0.6212
Family Income	0.2468	0.2158	0.2648	0.2332	-0.0605	0.1933
Single Parent Family	1.9637	0.8635	1.5698	0.9267	0.4292	0.8308
Home Province	1.5687	0.5961	1.4815	0.6050	0.4342	0.4297
Intercept	21.8445	6.5251	-20.2608	18.5802	62.0581	3.9181
2006 Freshman Cohort	4.9056	0.8118	1.6724	0.8918	4.2897	0.8074
2007 Freshman Cohort	2.5244	0.7197	0.2784	0.7483	2.8735	0.7694
Soph. RM Han Chin.	2.0370	2.9060	-6.3255	3.1550	10.4895	2.9922
Soph. RM Family Inc.	-0.7606	0.6061	-1.0970	0.5875	0.0843	0.5442
Soph. RM Single Parent	-0.6705	3.6971	1.5509	3.4310	-2.7334	3.8655
Soph. RM Home Prov.	1.3617	1.1083	-5.2320	2.0270	8.5792	1.1452
Fresh. RM Han Chin.	-0.7667	1.4203	0.6325	1.7212	-2.5432	1.4616
Fresh. RM Family Inc.	0.3488	0.3677	0.4009	0.3688	0.0495	0.3501
Fresh. RM Single Parent	1.0277	2.7186	-2.0290	2.2696	4.3731	2.0886
Fresh. RM Home Prov.	-0.6467	0.8602	0.7986	0.7969	-1.4838	0.6647
IV Han Chinese	-2.4004	0.9443	NA	NA	-1.9544	0.8817
IV Family Income	0.3314	0.3395	NA	NA	0.9757	0.3224
IV Single Parent Family	-1.7117	0.7965	NA	NA	-1.7678	0.7785
IV Home Province	-0.7783	1.5305	NA	NA	1.4478	1.3123
	Wald Test for IV		*J*-Test		Wald Test for IV	
Test Statistic (*p*-value)	10.6275 (0.0310)		3.9653 (0.2652)		18.0242 (0.0012)	

Notes: The full room sample of 464 observations is used for all the three regressions. The columns contain the estimates and standard errors for (in order from left to right) the indirect regression model of treated individual sophomore test scores, and two-step GMM causal model of treated individual sophomore test scores, and the first-stage model of the average sophomore test scores of the treatment individuals. The coefficients of the indirect regression and first-stage regression are estimated by OLS. All standard errors and the variance components in the test statistics are clustered at the room level.