Reconsidering Optimal Peer Assignment: How Does Non-Cognitive Peer Difference Matter?[*,†]

Li Li, Eric Mak, Chunchao Wang

February 18, 2019

Abstract

To facilitate knowledge spillover between students, the study of academic peer effects aims to guide the optimization of peer assignments (the seat arrangement of students) in classes. When brought to actual practice, a theoretically optimal assignment performed even worse than a randomized assignment. As we show, both theoretically and numerically, that successful knowledge spillover depends on non-cognitive peer difference (how peer personalities differ); its omission in the academic peer effect model specification generally leads to a sub-optimal peer assignment even in the presence of peer randomization. Quantitatively, the sub-optimal assignment amounts to a loss of academic achievement of 3% to 5% after a semester. We obtain this estimate by using data from a field experiment conducted in classrooms in China. We note that class teachers routinely readjust their peer assignments because assimilation—the convergence of non-cognitive peer difference over time—does not necessarily take place between deskmates.

Keywords: Peer Assignment; Academic Peer Effects; Non-Cognitive Peer Difference; Assimilation

[*] Li: lli.li@utoronto.ca; Mak: honam.mak@utoronto.ca; Wang: twangcc@jnu.edu.cn.

[†] All three authors contribute equally to this work. We sincerely thank Robert McMillan, Philip Oreopoulos, Shouyong Shi, and Aloysius Siow for their invaluable advice and insights. We also appreciate the seminar participants in University of Guelph, University of Toronto, Brock University, Haoren He, Fangwen Lu, and Le Wang for their comments. This paper was previously circulated under the title "Local Peer Group Optimization: Seat Rearrangements within Large Chinese Classrooms."

1 Introduction

The study of academic peer effects is based on a theory of knowledge spillover between peers (see Hoxby and Weingarth (2005); Sacerdote (2011) for extensive reviews). According to this theory, high-achieving students can potentially help low-achieving students when they are mixed in the same peer group. This theory leads to a policy problem of how to assign students into peer groups in a way that best exploits knowledge spillovers. Peer assignment across classes according to cognitive ability—ability tracking—is a central topic in education (Hoxby, 2000; Hoxby and Weingarth, 2005; Hanushek et al., 2003; Ding and Lehrer, 2007; Figlio, 2007; Aizer, 2008; Carrell and Hoekstra, 2010) despite its mixed results (Figlio and Page, 2002; Epple et al., 2002; Fu and Mehta, 2014). From a cost-benefit perspective, peer assignment is attractive because this education policy involves only the reallocation of existing resources—the students to be assigned—but not new investments. As another major education policy in discussion, reducing class size requires a significant investment in new teachers and classrooms and is hence more common among developed countries (Hanushek, 1998; Krueger, 1999, 2003; Chetty et al., 2011). As such, peer assignment could be particularly relevant when class size reduction is not feasible for financial or institutional reasons.

In a peer assignment problem, the policy-maker searches for the peer assignment that yields the highest welfare, which is generally defined as a weighted average of academic achievements among the students to be assigned.[1] Peer effect researchers contribute to solving the peer assignment problem by identifying the set of feasible outcomes that can be obtained by varying the peer assignment. For instance, mixing high- and low-achieving students would imply that high-achieving students would lose the chance of benefiting from their high-achieving peers. Determining the magnitude of this welfare loss requires the identification of academic peer effects. To this end, much progress has already been made through the use of (quasi-) experiments (Sacerdote, 2001; Carrell et al., 2009; Ammermueller and Pischke, 2009; Duflo et al., 2011; Booij et al., 2017). These experiments randomize peer assignment to eliminate the bias

[1] Similar to any welfare problem, some subjectivity is involved in deciding the welfare weights. For equality reasons, one may want to place a higher welfare weight on low-achieving students.

caused by peer selection.[2]

By now, a substantial body of experimental peer effect estimates is available to guide the solution of the peer assignment problem, although it was until Carrell et al. (2013) that first pointed out that the peer assignment problem is actually not as straightforward as it may seem. They assigned high-achieving students as peers of low-achieving students in a U.S. military college; following the request of the military, their objective was to best assist low-achieving students by manipulating the peer assignment. The feasible set of peer assignments was determined based on experimental peer effect estimates. However, against expectations, the theoretically optimized group performed even worse than the control group where peers were randomly assigned.[3]

Carrell et al. (2013) reported that the assigned peers may not necessarily get along with each other. We explore such argument by hypothesizing that the success of knowledge spillover does not simply depend on one's own and peer baseline scores, which proxy their respective cognitive abilities; given that its success also depends on how peer relationships manifest, such spillover depends on non-cognitive factors as well. We note that outside the peer effect context, education production functions (Krueger, 1999; Todd and Wolpin, 2003; Hanushek, 2008) have already incorporated non-cognitive inputs. The skill formation literature (Cunha and Heckman, 2007, 2008; Cunha et al., 2010) identifies a complementarity between cognitive and non-cognitive inputs. In parallel to the skill formation literature, we measure non-cognitive peer difference, defined as how one's own and his or her peer's Big Five measurements differ, and then introduce such difference into the standard academic peer effect model. Afterward, we explore the consequences of this extension to the peer assignment problem.

We raise and address the following questions:

[2] To be more specific, the idea behind peer randomization is that if peers are randomly assigned, then a comparison between the outcomes of two students whose baseline scores are the same would be entirely due to their respective peers than these students themselves being different in some way that is unobserved by the researcher.

[3] This leads to an important open question not necessarily restricted to education; the same applies to other peer contexts such as the workplace. This remark is true regardless of the data source, such as laboratory (Falk and Ichino, 2006), a particular workplace (a supermarket chain) (Mas and Moretti, 2009) or from representative data (Cornelissen et al., 2017).

1. If the policy-maker can assign peers only according to baseline scores, would he obtain a sub-optimal assignment because the non-cognitive peer difference is ignored?

2. Quantitatively, can the policy-maker do significantly better if she can assign peers according to non-cognitive peer difference as well?

3. Do policy-makers assign peers according to non-cognitive peer difference, which suggests that the previous literature has assumed a smaller set of feasible assignments than actual practice allows?

The answers to these three questions are all affirmative.

To address question 1, we show that randomized peer assignment does not necessarily guarantee, as typically assumed, the identification of peer effects. As we show, peer randomization only eliminates the selection bias (caused by own unobservables) but not the bias caused by peer unobservables. Most studies that apply peer randomization since Sacerdote (2001) interpret peer unobservables as classical measurement error.[4] If the error term contains a true non-cognitive input not measured by the researcher, then generally this term biases the peer effect estimates. We then consider a bilateral assignment example, where the optimal assignment should be positive assortative matching with respect to the peer effects. We show how this new bias generally change the optimal matching.

To address question 2, we estimate our extended peer effect model. Our estimation results find that non-cognitive peer difference is critical, that is, having high-achievers as peers leads to a positive academic peer effect (an elasticity of about 0.1) if and only if non-cognitive peer difference is relatively small (1 SD below mean). Meanwhile, a reverse-signed academic peer effect with a comparable magnitude is found if the non-cognitive peer difference is relatively large (1 SD above mean). Such interaction between baseline scores and non-

[4] Since Carrell et al. (2013) reported their surprising results, the subsequent literature has provided some discussion on the effects caused by measurement error. Angrist (2014) shows that when both own and peer observables (the baseline scores) are subject to measurement error, peer effect estimates can be biased in any direction; this bias may misguide the optimization step in Carrell et al. (2013). However, in case the peer assignment is random, Feld and Zölitz (2017) show that only standard attenuation bias remains.

cognitive peer difference matters in the subsequent peer assignment problem. Without the interaction term, the optimization result based on the confounded peer effect estimates is indistinguishable from a random assignment benchmark, which is similar to the result in Carrell et al. (2013). With the interaction term, optimization yields a 3% to 5% improvement for all classes within one semester, depending on the subject.

To address question 3, we estimate our model using data from a field experiment conducted in Chinese primary classrooms because of practical relevance. The lack of education resources and the rapid urbanization in sub-urban China result in very large class sizes that typically includes over 40 students per class (see Section 2 for details). In this context where class size reduction is difficult to achieve, Chinese classrooms are managed by having the seats centrally assigned by a "class teacher." In solving the peer assignment problem, class teachers in China make frequent adjustments to their seating plans whenever peer conflicts arise. From our data we find that assimilation (the convergence of non-cognitive peer difference over time) fails with about probability a half, and that assimilation failure leads to a large (17%) reduction in friendships (in terms of probability). Once we take assimilation into account, holding baseline scores fixed, the corresponding optimal peer assignment is almost entirely changed. Our notion of assimilation operates between peers as in psychology (Harris, 1995) and is similar to macroscopic assimilation concerning immigrants (Chiswick, 1978; Borjas, 1985, 1995, 2015; Lazear, 1999; Meng and Gregory, 2005; Konya, 2007; Abramitzky et al., 2016) who cannot avoid the locals.

We find aggregate assimilation patterns that are consistent with those found in the existing literature. Specifically, we find that female-female pairs assimilate much better than other possible gender pairs, thereby supporting the results of another within-classroom peer effect study (Lu and Anderson, 2015) where female-female pairs are found to have positive academic peer effects within classes.[5] We also find that younger, lower grade peers have better as-

[5] However, we find a strong reverse gender gap (Goldin et al., 2006; Fortin et al., 2015) such that females are mostly high-achieving and that males are mostly low-achieving; consequently, we have few opportunities to form female-female pairs in our peer assignment. Notably, our conclusion cannot be drawn by just examining the peer effect estimates without evaluating the peer assignment problem.

similation than their older, higher grade counterparts. This result agrees with another major conclusion from the skill formation literature, that is, early education inputs yield large returns.

As a within-classroom peer effect study, our work guides the assignment of seats within classes while holding class composition constant rather than investigating how class composition affects a particular student within the class.[6] In many countries, ability tracking is discouraged for equity reasons. For the case of China, this practice is explicitly banned by law for compulsory education.[7] Relatively speaking, seat assignment within a class is easier to implement than ability tracking, both practically and politically.

We conduct several robustness checks and obtain the following results: 1) peer effects are local, operating within current deskmates only; and 2) our results are robust to the choice of non-cognitive measures in measuring non-cognitive peer difference. In the appendix, we discuss several issues including measurement errors, wild bootstrap clustered standard errors, and the context of large Chinese classrooms, in detail.

The rest of this paper is organized as follows. Section 2 describes the practical context of peer assignments within Chinese classrooms, our field experiment design, and the measurement of non-cognitive peer difference. Section 3 describes our peer effect model and explains our econometric arguments in detail. Section 4 reports the academic peer effect estimates. Section 5 discusses the results of our optimization attempt. Section 6 reports the assimilation patterns. Section 7 concludes our paper.

2 Context, Field Experiment, and Data Description

We choose to study academic peer effects in classrooms in China. China ranks first worldwide in average class size— a clear outlier with 37 students per class relative to the OECD average of 21. The typical class size in many schools

[6] Several macroscopic classroom peer effect papers have discussed non-cognitive related issues such as class disruptions (Lazear, 2001); see Sacerdote (2011, 2014) for comprehensive reviews.

[7] See Article 22, 57 of the Education Law, People's Republic of China.

located in the sub-urban areas of China is above 40. In our sample (to be described in detail below) the class size ranges from 37 to 64. These figures are beyond the class size studied in the previous class size literature.[8] The major theoretical justification behind class size reduction is the “Bad Apple Principle" (Lazear, 2001), where disruptions due to peer conflicts increase along with class size. Accordingly, China’s large class size should be difficult to sustain.

Education researchers have attributed the large classes in Chinese primary schools to several reasons (Jin and Cortazzi, 1998; Teng and Liang, 2012). First, due to the rapid urbanization in China, most rural children tend to live in county towns (mostly the county capitals) instead of their surrounding villages.[9] Second, local governments have been consolidating education resources by closing down village primary schools. Third, parents are aware of the better education quality of primary schools located in the county towns, thereby sending their children to these county primary schools even if a village primary school is located nearby. Given these push and pull factors, the primary schools located in county towns are typically oversized. The class size in some locations are surprising. In a provincial study of the education system in Henan Province (adjacent to Hubei Province in our study), Teng and Liang (2012) found that the class size in a "key-point" (elite) primary school amounted to 133 students in 2011.

Instead of reducing class size, China solves the class management problem through the so-called "class teacher system." A class teacher, who is also the teacher of a particular subject, specializes in managing a class and she is responsible to design the seating plan.[10] While the actual practice varies, a class teacher typically reassigns the seats within her class whenever she observes serious conflicts between a pair of deskmates. Furthermore, a class teacher

[8] For instance, the well-known STAR program reduces the class size from the standard size of 22 to a smaller size of 15. To provide additional references for class size, Angrist and Lavy (1999) adopted a regression continuity design to study the effect of class size with the threshold being 40 students; the cutoff in a similar Swedish study (Fredriksson et al., 2012) is 30. Among developing countries, Urquiola (2006) examined the case of Bolivia, where the class size is around 30 to 40 students.

[9] Since 1978, urban population has increased from 18% to almost 60% by 2015, the period of our study. See Henderson et al. (2009) for a recent general review of China’s urbanization.

[10] Given the lack of a direct English equivalent of the concept of class teacher, we adopt our own translation.

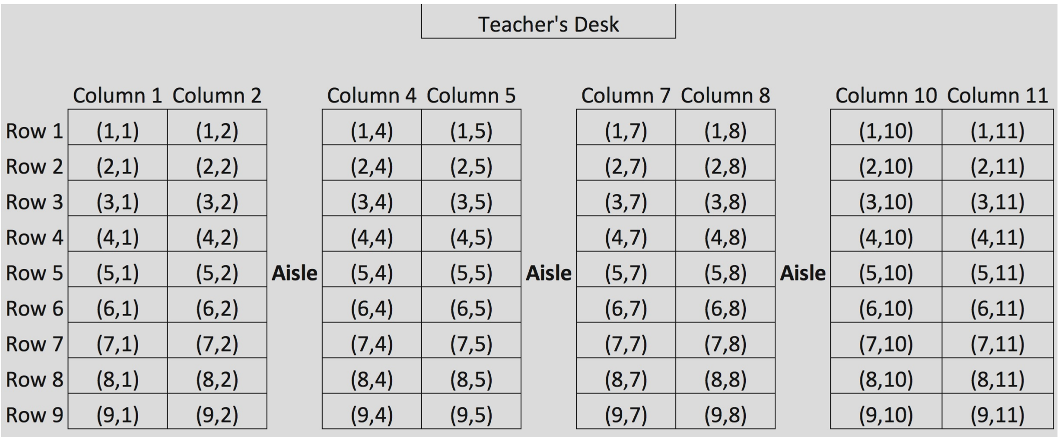
would base on other criteria, such as gender (mixed gender is often preferred), to decide who should seat next to whom. Such adjustment to the peer assignment continues throughout the semester. This continual adjustment of the peer assignment is fundamentally different to Carrell et al. (2013) and the academic peer effect literature in general, which presumes the existence of an optimal time-invariant peer assignment and searches for it.

As we observe, the class teacher system—a nationalwide instance of peer assignment in actual practice—naturally emerges due to cost-benefit concerns. As discussed in the Introduction, peer assignment costs little to implement since it involves only the reallocation of existing resources. By contrast, reducing class size (Hanushek, 1998; Krueger, 1999; Chetty et al., 2011) involves the employment of new teachers and increasing the number of classrooms. Reducing class size is particularly difficult in remote rural school districts of China where attracting a sufficient number of qualified teachers from cities is already difficult. A recent policy discussion of reducing class size in rural China has been concerned to reduce class sizes of over 100 to a more reasonable number, say, 60. The class teacher system is arguably a viable alternative, and has been effectively applied to maintain classroom order in China for several decades.

To provide the reader with a general idea of how this system operates, Figure 1 shows an example seating plan within a typical class in China. Each classroom is organized as an rectangular array. A rectangular array arrangement fits in the largest number of students within a classroom, and is the standard seating arrangement found in most Chinese classrooms. To minimize confusions in seating, each student is assigned a fixed seat. For easy access to every seat in the classroom, every two columns is separated by an aisle. As such, students sit in pairs, and each student has a fixed deskmate unless reassigned by the class teacher.

In our sample of 21 classes in three schools in Hubei, China, we record the x-y coordinates of the seats for the purpose of identifying deskmates. We identify the deskmates of 892 out of 1005 students. The remaining students sit alone so that they do not have deskmates.[11] The classrooms in our sample vary

[11] For example, the student seating in the second row, fourth column has the coordinates $(2, 4)$ in our data-set, while his deskmate has the coordinates $(2, 5)$. One aisle occupies a column (e.g., the leftmost aisle after the first and second rows is labelled column 3 (missing)). This setup yields the following simple algorithm to identifying



Teacher's Desk

	Column 1	Column 2		Column 4	Column 5		Column 7	Column 8		Column 10	Column 11
Row 1	(1,1)	(1,2)		(1,4)	(1,5)		(1,7)	(1,8)		(1,10)	(1,11)
Row 2	(2,1)	(2,2)		(2,4)	(2,5)		(2,7)	(2,8)		(2,10)	(2,11)
Row 3	(3,1)	(3,2)		(3,4)	(3,5)		(3,7)	(3,8)		(3,10)	(3,11)
Row 4	(4,1)	(4,2)		(4,4)	(4,5)		(4,7)	(4,8)		(4,10)	(4,11)
Row 5	(5,1)	(5,2)	**Aisle**	(5,4)	(5,5)	**Aisle**	(5,7)	(5,8)	**Aisle**	(5,10)	(5,11)
Row 6	(6,1)	(6,2)		(6,4)	(6,5)		(6,7)	(6,8)		(6,10)	(6,11)
Row 7	(7,1)	(7,2)		(7,4)	(7,5)		(7,7)	(7,8)		(7,10)	(7,11)
Row 8	(8,1)	(8,2)		(8,4)	(8,5)		(8,7)	(8,8)		(8,10)	(8,11)
Row 9	(9,1)	(9,2)		(9,4)	(9,5)		(9,7)	(9,8)		(9,10)	(9,11)

Figure 1: A Sample Seating Plan

in their dimensions, with some having more rows than the others. As such, the actual classrooms are typically slightly smaller than the rectangular 9×11 array shown in Figure 1, although not by much (the average is 7×11).

Table 1 presents the class-level summary statistics of our sample. One notable feature is class size: the smallest class has a class size of 37, while the largest class has a size of 64; the mean class size is 47.9. The typical number of rows and columns being 7 and 11 (including 3 aisles), so that a considerable physical distance is observed between the teacher and the furthermost students.

While the actual practice varies, from our discussion with the class teachers, we identify several common patterns of peer assignment:

1. The class teacher would typically assign students according to their height, so that shorter students sit in the front rows. Height as an assignment critierion is necessary because one must minimize the possibility of having front row students blocking the vision of back row students. A substantial within-class variation in height can be observed in our sample (standard deviation being 7.7 cm, with respect to a mean of 137 cm).

the deskmate of student i:

1. The deskmate shares the same row number $x[i]$ as i.
2. The deskmate is the neighbor in columns, i.e. the mapping is $1 \to 2, 2 \to 1, 4 \to 5, 5 \to 4, 7 \to 8, 8 \to 7, 10 \to 11, 11 \to 10$, with columns $3, 6, 9$ as aisles.

Table 1: Class-Level Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Year of Birth (Class Mean)	21	2007	0.683	2005	2007
Year of Birth (Class Std. Dev.)	21	0.531	0.106	0.380	0.776
Baseline Height (Class Mean)	21	137.400	3.706	130.100	143.800
Baseline Height (Class Std. Dev.)	21	7.658	2.005	3.126	11.470
Proportion of Male	21	0.560	0.062	0.451	0.692
Class Size	21	47.860	6.850	37	64
Number of Rows	21	7.190	1.030	6	9
Number of Columns (including aisles)	21	11.000	0.000	11	11

2. Class teachers have a strong tendency to create pairs with mixed gender. The general belief is that mixed-gender pairs are relatively stable.

3. Initially, the class teacher cannot easily identify whether deskmates can get along with each other. Therefore, the initial seating plan is often based on the plan used in the previous academic year; otherwise, the students are randomly assigned. Subsequently, the class teacher rearranges the seats in a trial and error manner, depending on his or her observation on how each student behaves.

4. Those seats located close to the teacher's desk are generally preferred by students. Therefore, parents usually ask teachers to have their children seated in these preferred seats. This finding implies the importance of a student's location in the classroom in general. Consequently, we control for the exact location in the class by using fixed effects in our regressions.

5. To even out the disadvantages of some seats (some columns are located further away from the teacher's desk), some teachers rotate the seats periodically by rotating columns in pairs. As such, the deskmates pairs are preserved. We do not regard these rotations as seat reassignments in this paper.

The three primary schools in our sample have similar daily schedules, starting early at around 7am with morning recitations, followed by a breakfast at school, and then by regular classes. The first three 40-minute regular classes take place between 9AM to 11:30AM with a break in between. After lunch, the fourth and fifth class take place between 2PM to 4PM. The next 45 minutes remaining time is allocated for extra-curricular activities. During these five regular classes, students stay in their seats and are strictly forbidden to walk around the classroom. Since most students live far away from the school, they usually eat their breakfast and lunch in the classrooms at their own seats. In this environment, a student would interact with his or her deskmate for a substantial amount of time per day.

In our field experiment, we requested the three schools to randomize the seat arrangements for all their Grade 3 to 5 classes. To minimize the possibility of having taller students blocking the sight of shorter students, the randomization of seats is conditional on height. Our randomization scheme is similar to that of Lu and Anderson (2015).[12] However, unlike Lu and Anderson (2015), we enforce the randomization to the entire class without allowing for preferential treatments. Being aware that teachers and parents might have a tendency to overrule our randomization, we strictly monitored the classes. Our monitoring team regularly invigilated these schools.[13]

Concerning data collection, the three schools provided us with the baseline, mid-term and the final scores for each semester. We focus on Chinese Language and Math because for most (3/4) of the sample, the English baseline score is unavailable. We also collected supplementary information about students' self-reported friendships, and own and peer non-cognitive skills in our survey. Specifically, we use the Big Five, a widely-used taxonomy that encompasses most of the relevant psychological traits (Costa Jr and McCrae, 1992). The Big Five is measured by a vector of Likert-style questionnaire items (an ordinal score in the range 1-5). In this paper, we adopt a Chinese version of the Revised NEO Five-Factor Inventory (NEO-FFI) (Costa and MacCrae, 1992),

[12] At the beginning of the first semester, students are sorted by height, then grouped into four blocks. Within each block, students have their seats assigned randomly by a lottery. After the students in one block are randomly assigned to two rows following this procedure, the students in next block are assigned to the next two rows and so on. In this way, the students in different blocks will not be seated in the same row.

[13] To this end, we formed a monitoring team consisted of graduate students from Jinan University, China, and staff from the Hubei Education Bureau. As an administrative order, the teachers in each class are asked to provide the finalized seating plan and take photographs each week during the semester, making sure that the seating plan was followed through. Due to confidentiality, we do not include these photos in this paper but they are available upon request.

which consists of 60 items selected from a longer version of the questionnaire for their strong correlations with their associated factor scores.

3 Peer Effect Model

3.1 The Standard Academic Peer Effect Model

We first show that peer randomization would eliminate the selection of peers, but it cannot distinguish observable and unobservable peer effects from each other. Formally, let $i \in \mathbb{S} = \{1, 2, \ldots, N\}$ index individuals. $t = 0$ is the baseline period where peer randomization takes place, and $t = 1$ is the time where the current academic achievement is measured. We denote the academic achievement for individual i at time t by y_{it}.

A classroom with fixed seats defines a known peer function $p : \mathbb{S} \to \mathbb{S}$ such that $p(i) \in \mathbb{S}$ indicates the deskmate of individual $i \in \mathbb{S}$. We write $y_{p(i)0}$ to represent the baseline test score achievement of $p(i)$.[14]

A standard academic peer effect model considered in Carrell et al. (2013) can be expressed as follows:

$$y_{i1} = f(y_{i0}, y_{p(i)0}) + u_{i0} + \lambda u_{p(i)0} \tag{1}$$

where u_{i0} is the own unobservable error term, and $u_{p(i)0}$ is the peer counterpart. $f(.)$ is a non-linear function that depends on own and peer baseline score. $\lambda \in \mathbb{R}$ is a parameter capturing the strength of unobservable peer influences. Note that the peer effect model (1) is symmetric in the treatment of observable and unobservable inputs, in the sense that for both kinds of input, the own and peer components are explicitly stated. In most peer effect models, the unobservable peer inputs are subsumed into a correlated effect— a common factor shared by both i and $p(i)$ (see Manski (1993)); here, we introduce u_{i0} and $u_{p(i)0}$ as two separate unobservable inputs. The case with common unobservable inputs between own and peer is a sub-case of ours.

Note also that most of the peer effect literature discusses a static setup. In these static peer effect models, the concurrent influence of peer outcomes on

[14] If p represents deskmates, then $p(p(i)) = i$ so that $p \circ p$ is the identity mapping.

own outcomes leads to the classic reflection problem (Manski, 1993). Nevertheless, in many academic peer effect papers, e.g. Sacerdote (2001), data on baseline score are available. Consequently, several academic peer effect papers directly consider a one-period lag specification where the main independent variable is peer baseline score instead of peer current academic performance. In this case, the reflection problem does not apply. We follow the same approach in this work. In the context of academic performance, a strictly simultaneous peer effect on current academic achievement—the exam scores—cannot exist, under the assumption that no cheating has taken place during the exams.

Peer randomization is important in the estimation of causal peer effects. Peer randomization at $t = 0$ implies that for any measurement, X_{i0} and $X_{p(i)0}$ are i.i.d. draws from the same underlying population X, whether X is observable or not. Therefore,

$$(y_{i0}, u_{i0}) \perp\!\!\!\perp (y_{p(i)0}, u_{p(i)0}) \tag{2}$$

Conceptually, consider the following group comparison:

$$E[y_{i1}|y_{i0} = y^*, y_{p(i)0} = y_H] - E[y_{i1}|y_{i0} = y^*, y_{p(i)0} = y_L] \tag{3}$$

which is to compare, among students having the same own baseline score y^*, the current academic achievement y_{i1} of two comparison groups, namely, 1) those who have peers with a higher baseline score y_H, and 2) those who have peers with a lower baseline score y_L. According to (1), this between-group comparison groups identifies the following:

$$\begin{aligned}
&f(y^*, y_H) - f(y^*, y_L) \\
&+ E[u_{i0}|y_{i0} = y^*, y_{p(i)0} = y_H] - E[u_{i0}|y_{i0} = y^*, y_{p(i)0} = y_L] \\
&+ \lambda E[u_{p(i)0}|y_{i0} = y^*, y_{p(i)0} = y_H] - \lambda E[u_{p(i)0}|y_{i0} = y^*, y_{p(i)0} = y_L]
\end{aligned} \tag{4}$$

$f(y^*, y_H) - f(y^*, y_L)$ is the marginal effect of changing the peer baseline score, ceteris paribus. The sum of the two terms in the second line is zero because the own unobservable inputs for the two comparisons groups are balanced due to peer randomization at $t = 0$, which eliminates the selection effect. However, given that the average peer unobservable inputs are not balanced by the peer randomization, the sum of the terms in the third line is generally non-zero.

We now discuss the consequences of this confounding in the peer assignment problem. We follow Carrell et al. (2013) by assuming the objective function is to maximize the average current achievement score among the low-achieving students. Specifically, we consider the following optimization problem. Let $\mathbb{S}_{low} \subset \mathbb{S}$ be a set of students whose current academic achievement is to be maximized in expected terms, and let $\mathbb{S}_{high} = \mathbb{S} \setminus \mathbb{S}_{low}$ be its complement. $\mathbb{S}_{low}$ is defined as the set of students whose baseline score is below the median. For simplicity, suppose that n is even so that $\mathbb{S}_{low}$ and $\mathbb{S}_{high}$ are equal sized. As such, the problem becomes one that searches for a mapping $q : \mathbb{S}_{low} \to \mathbb{S}_{high}$ that maps a deskmate from $\mathbb{S}_{high}$ to each student from $\mathbb{S}_{low}$. We denote the corresponding space of feasible mappings by $\mathcal{Q}$.

The standard academic peer effect model leads to the following peer assignment problem:

$$\max_{q \in \mathcal{Q}} \frac{1}{n/2} \sum_{i \in \mathbb{S}_{low}} f(y_{i0}, y_{q(i)0}) \tag{5}$$

Note that f must be non-linear, otherwise the peer assignment problem is indeterminate because all assignments $q \in \mathcal{Q}$ would lead to the same objective value.

The ideal peer assignment problem should be one that incorporate the own and peer unobservables, that is:

$$\max_{q \in \mathcal{Q}} \frac{1}{n/2} \sum_{i \in \mathbb{S}_{low}} [f(y_{i0}, y_{q(i)0}) + u_{i0} + \lambda u_{q(i)0}] \tag{6}$$

However, given that the unobservable component $\sum_{i \in \mathbb{S}_{low}} [u_{i0} + \lambda u_{q(i)0}]$ is invariant to the choice of q, it can be eliminated from the objective function, resulting in (5). This result critically relies on the linearity of the unobservable component.

As we emphasize, a more fundamental concern in using (5) as the objective function is that f cannot be identified due to the confounding problem, yet the optimization algorithm assumes so. To illustrate the consequences of this problem, consider an example where

$$f(y_{i0}, y_{q(i)0}) = \beta_0(y_{i0}) + \beta_1(y_{i0}) y_{q(i)0}$$

which includes a typical specification in academic peer effect regressions where y_{i0} interacts multiplicatively with $y_{p(i)0}$. In this example, the estimated marginal effect of raising peer baseline score, holding own baseline score fixed at $y_{i0} = y^*$ can be expressed as

$$\tilde{\beta}_1(y^*) \equiv \beta_1(y^*) + \lambda \frac{\partial \mathbb{E}[u_{p(i)0}|y_{i0} = y^*, y_p]}{\partial y_p}$$

instead of $\beta_1(y^*)$.

Notice that the assumed functional form is supermodular with respect to $\beta(y_{i0})$ and $y_{p(i)0}$. Therefore, the true optimal assignment would be positive assortative matching (PAM), i.e. suppose that for $i, j \in \mathbb{S}_{low}$ such that if $\beta_1(y_{i0}) > \beta_1(y_{j0})$, then $y_{q(i)0} > y_{q(j)0}$. However, given that $\tilde{\beta}_1(y^*)$ is being identifed instead, the computed optimal assignment would generally be different from the true optimal one unless a strictly increasing function g exists such that

$$g(\beta_1(y^*)) = \tilde{\beta}_1(y^*)$$

for all y^*, that is, the observable and unobservable peer effects are ordered in the same way. This monotonicity condition is less demanding than exogeneity, i.e. shutting down unobservable peer effects:

$$\lambda \frac{\partial \mathbb{E}[u_{p(i)0}|y_{i0} = y^*, y_p]}{\partial y_p} = 0$$

While weaker than exogeneity, the monotonicity condition has no guarantee to be satisfied in the data, thereby likely resulting in a sub-optimal peer assignment.

3.2 Introducing Non-Cognitive Peer Difference

We now study the case where the policy-maker can perform peer assignments based on non-cognitive peer difference, which is previously assumed to be unobservable. The own and peer non-cognitive skills at time 0 are denoted by $n_{i0}, n_{p(i)0}$ respectively. In our work, they are measured by our Big Five measurements. We define a metric of non-cognitive peer difference:

$$d_{i0} \equiv |n_{i0} - n_{p(i)0}|$$

The resulting peer effect model is:

$$y_{i1} = f(y_{i0}, y_{p(i)0}, d_{i0}) + u_{i0} + \lambda u_{p(i)0} \tag{7}$$

$$d_{i1} = g(y_{i0}, y_{p(i)0}, d_{i0}) + v_{i0} + \gamma v_{p(i)0} \tag{8}$$

which resembles the skill formation model (Cunha and Heckman, 2007, 2008; Cunha et al., 2010). In the skill formation model, the inputs are own cognitive and non-cognitive skills in the previous period. Our peer effect model introduces the peer counterparts.

Some discussion of our choice of functional form is warranted. We construct the metric d_{i0} rather than letting $n_{i0}, n_{p(i)0}$ to enter freely in (7) and (8). This restriction is due to our notion of non-cognitive peer difference. The Big Five, while being comprehensive, does not yield a univariate ranking. People are generally are "different from" rather than "better/worse" under the Big Five since it involves more than one dimension. In this paper, we are particularly interested in how peer non-cognitive differences affect academic achievement. Therefore, for our purpose, only a metric that measures the non-cognitive differences between a pair of deskmates, rather than their respective levels, is needed.

Given that the identification analysis of this model is similar to the above, we do not repeat the analysis. Extending the model to include general neighbor(s) is straightforward. Let $n(i)$ denote another peer of individual i.[15] Peer randomization at time 0 guarantees that both $p(i)$ and $n(i)$ are randomly assigned, such that:

$$(y_{i0}, d_{i0}, u_{i0}, v_{i0}) \perp\!\!\!\perp (y_{p(i)0}, d_{p(i)0}, u_{p(i)0}, v_{p(i)0}) \perp\!\!\!\perp (y_{n(i)0}, d_{n(i)0}, u_{n(i)0}, v_{n(i)0}) \tag{9}$$

holds, i.e. the own, deskmate, and general neighbor variables are independent

[15] In the more general case with multiple neighbors, then we can extend the definition of p to be a set function; we consider the "representative peer" so that for each individual i and any measurement Z, $Z_{p(i)t}$ is its average over all neighbors in $p(i) \subseteq \mathbb{S}$ at time t.

to each other, whether observable or not. The following analysis is analogous to the above. To keep the notation simple, we continue our discussion while suppressing general neighbor terms.

Analogous to the above, we consider the following optimization problem:

$$\max_{q \in \mathbb{Q}} \frac{1}{n/2} \sum_{i \in \mathbb{S}_{low}} [f(y_{i0}, y_{q(i)0}, |n_{i0} - n_{q(i)0}|)] \tag{10}$$

which requires $n_{i0}, n_{q(i)0}$ to be observable. Notably, the non-cognitive peer difference enters f in a non-linear manner. We consider an example:

$$f(y_{i0}, y_{q(i)0}, |n_{i0} - n_{q(i)0}|) \equiv \beta(y_{i0}) y_{q(i)0} \kappa(|n_{i0} - n_{q(i)0}|)$$

Considering non-cognitive peer difference can break down the PAM assignment because changing the assignment $q(i)$ would affect both $y_{q(i)0}$ and $|n_{i0} - n_{q(i)0}|$. In this case, the latter "matching effect," if dominating, can change the optimal assignment pattern. Whether the two effects offset or reinforce each other depends on two factors: 1) the joint distribution of peer baseline score and non-cognitive peer differences and 2) how non-cognitive peer difference affects the total peer effect as determined by the function $\kappa(.)$. How the optimal assignment would change is a numerical question that we explore in Section 5.

4 Empirical Results

4.1 Academic Peer Effect Estimates

We report our major peer effect regression results in this section. We run least square regressions of Chinese Language test scores on the own Chinese Language baseline score, deskmate baseline score, and baseline non-cognitive peer difference d_{i0}. We take logs of all test scores, so that the peer effect estimates are elasticities. We also run the corresponding regressions for Math.

For presentation purposes, in this set of regressions, we standardize d_{i0} to have a zero mean and unit variance. For the same reason, we also divide the whole sample into two groups according to own baseline score and then run the peer effect regressions separately for each sub-group.

We are also interested in comparing deskmates and more general neighbors. Therefore, following Lu and Anderson (2015) and Hong and Lee (2017), we construct the neighbor-4 measure corresponding to each variable, including the two neighbors in the front and the two neighbors at the back. For those students seated in the first row or the last row, some neighbors could be missing. Correspondingly, we define the neighbor measures by averaging over the available neighbors. In this set of regressions, all non-cognitive peer difference variables are standardized to have a zero mean and unit variance.

In the peer effect regressions, we control for several confounding factors:

1. First, we control for class fixed effects to capture the confounding variations related to between class heterogeneity.

2. Second, given the large size of classrooms, the locations in the classroom matter; therefore, we control for the exact location fixed effects (a dummy for each $x-y$ coordinate). This set of controls is finer than only controlling for the randomization block (the quartiles in our experiment) as in Lu and Anderson (2015).

3. Third, we control for gender pairs. Specifically, for own-deskmate pairs, we introduce male-female, female-male, and female-female dummies. For neighbors, we introduce dummies to capture all possible variation in own-neighbor gender pairs; since there are four neighbors, the neighbor gender is a simple average over them (e.g., 0.75 if 3 out of 4 of them are male).

The results for Chinese and Math are presented in Tables 2 and 3, respectively. In each table, we present three variants of the peer effect regressions that use the whole sample, the observations with own baseline score below and above the median. The purpose of dividing the sample according to own baseline score is to explore non-linearity while avoiding a complicated discussion of the three-way interactions between own and peer baseline scores and non-cognitive peer difference. In our peer assignment optimization, we consider a full three-way interaction.

Several features emerge from these peer effect estimates. If the own baseline score is above the median, then there is no peer effect and only the low-achievers can potentially benefit from having a peer with a high baseline score.

Consequently, we can focus on the welfare of low-achieving students. We also find that the peer baseline score coefficients are almost zero (and statistically insignificant), while the interaction term between peer baseline score and non-cognitive peer difference matters. For example, take the Chinese Final interaction coefficient of -0.104 (for students with own baseline score below the median). This coefficient implies that, the elasticity of peer baseline score for low-achieving students with respect to Chinese final is:

- approximately 0.10 if the non-cognitive peer difference is $-1s.d.$ relative to its mean (close peer hereafter).
- ~ 0 if the non-cognitive peer difference is at its mean; and
- -0.10 if the non-cognitive peer difference is $+1s.d.$ relative to its mean (distant peer hereinafter).

A similar intepretation can be applied for other interaction coefficients in the tables. To conclude, we find that assigning a high achiever as one's peer has opposite signed effects, depending on whether the peer is a close peer or a distant peer. That is, if a high-achieving peer is also a close peer, he benefits own academic achievement; otherwise, he could possibly reduce his or her own academic achievement.

Our results relate to some well-known classroom peer models (see Hoxby and Weingarth (2005) for a comprehensive review). Among them, the "Boutique model" proposes that having homogeneous peers are beneficial to own achievement, while the "Shining Light" model proposes that high-achieving students set good examples to their low-achieving peers. Generally, the set of sensible scenarios of peer interactions (that could be justified by theoretical models) is rather large, and empirical studies must be performed to find out which among these scenarios are true. The situation could grow more complicated when more than one dimension—baseline score and non-cognitive peer difference—needs to be considered.

We find that the best combination is to have a high-achieving student as a peer, given that he is close relative to own in terms of non-cognitive skill; as such, our results are consistent with the Boutique model along the non-cognitive dimension, and the Shining Light Model along the cognitive dimension. Furthermore, a high-achieving peer who is also a distant peer can potentially discourage own achievement, consistent with the "Invidious Comparison" model.

The aforementioned classroom peer effect models in Hoxby and Weingarth (2005) concern global peer effects, whereas our results focus on peer effects that are local between deskmates only. Specifically, we do not find any statistically significant result for general neighbors, whose peer effect estimates are an order of magnitude smaller than the corresponding ones for deskmates. Our placebo test in the Appendix demonstrates the same conclusion.

Table 2: Peer Effect (Chinese)

	First Sem. Midterm			First Sem. Final		
	(1)	(2)	(3)	(4)	(5)	(6)
Own Baseline	0.771***	0.763***	0.780***	0.592***	0.556***	0.844***
	(0.047)	(0.049)	(0.071)	(0.056)	(0.059)	(0.099)
Deskmate1 Baseline	−0.032	−0.046	−0.018	0.018	−0.016	0.015
	(0.028)	(0.034)	(0.013)	(0.019)	(0.018)	(0.018)
Non-Cog Diff (Deskmate1)	−0.006	0.003	−0.001	0.298***	0.392***	−0.010
	(0.090)	(0.101)	(0.034)	(0.060)	(0.104)	(0.039)
Neighbor1 Baseline	−0.034	−0.118	−0.022	−0.044	−0.066	−0.020
	(0.030)	(0.095)	(0.019)	(0.063)	(0.134)	(0.019)
Non-Cog Diff (Neighbor1)	−0.035	−0.149	−0.020	−0.155	−0.162	−0.066
	(0.164)	(0.540)	(0.062)	(0.103)	(0.207)	(0.055)
Interaction (Deskmate1)	−0.005	−0.016	−0.001	−0.074***	−0.104***	0.001
	(0.024)	(0.031)	(0.008)	(0.017)	(0.032)	(0.009)
Interaction (Neighbor1)	0.015	0.048	0.006	0.042*	0.051	0.015
	(0.042)	(0.134)	(0.014)	(0.024)	(0.050)	(0.012)
class FE	Yes	Yes	Yes	Yes	Yes	Yes
location FE	Yes	Yes	Yes	Yes	Yes	Yes
gender pair FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	882	444	452	879	442	451
R^2	0.636	0.653	0.495	0.697	0.727	0.551
Adjusted R^2	0.582	0.541	0.332	0.652	0.638	0.406

Notes:

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.
Standard errors are obtained by wild bootstrap and clustered at class level.
All test scores are in logs.
(1),(4): Whole sample
(2),(5): Own baseline below the median
(3),(6): Own baseline above the median
Baseline non-cognitive peer differences are standardized.

Table 3: Peer Effect (Math)

	First Sem. Midterm			First Sem. Final		
	(1)	(2)	(3)	(4)	(5)	(6)
Own Baseline	0.843***	0.819***	1.215***	0.845***	0.884***	0.955***
	(0.060)	(0.076)	(0.079)	(0.079)	(0.079)	(0.080)
Deskmate1 Baseline	−0.018	−0.043	−0.038*	0.062	0.069	−0.029**
	(0.030)	(0.040)	(0.022)	(0.047)	(0.070)	(0.015)
Non-Cog Diff (Deskmate1)	0.121	0.328***	−0.076	0.393*	0.673**	−0.036
	(0.111)	(0.119)	(0.082)	(0.215)	(0.297)	(0.050)
Neighbor1 Baseline	0.014	−0.007	−0.023	0.086	0.163	0.030
	(0.039)	(0.054)	(0.045)	(0.063)	(0.116)	(0.036)
Non-Cog Diff (Neighbor1)	0.107	−0.125	0.240*	−0.155	−0.336	0.190
	(0.081)	(0.177)	(0.130)	(0.217)	(0.304)	(0.180)
Interaction (Deskmate1)	−0.023	−0.066**	0.019	−0.089*	−0.153**	0.009
	(0.025)	(0.027)	(0.019)	(0.048)	(0.067)	(0.011)
Interaction (Neighbor1)	−0.028	0.023	−0.055*	0.038	0.081	−0.042
	(0.019)	(0.042)	(0.029)	(0.050)	(0.070)	(0.041)
class FE	Yes	Yes	Yes	Yes	Yes	Yes
location FE	Yes	Yes	Yes	Yes	Yes	Yes
gender pair FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	882	465	442	879	463	441
R^2	0.761	0.787	0.554	0.642	0.676	0.476
Adjusted R^2	0.726	0.722	0.406	0.589	0.576	0.301

Notes:
***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.
Standard errors are obtained by wild bootstrap and clustered at class level.
All test scores are in logs.
(1),(4): Whole sample
(2),(5): Own baseline below the median
(3),(6): Own baseline above the median
Baseline non-cognitive peer differences are standardized.

5 Peer Assignment Optimization

Given the causal estimates reported in Section 4, we are now in a position to discuss the implementation details of the peer assignment optimization and report the associated results. As a computational note, the size of the set of feasible assignments $\mathcal{Q}$ is $(n/2)!$. Given the large class sizes in our sample, brute force computation is not feasible. As one well-known way of attack, we relax the problem by allowing for fractional assignment, thereby transforming the problem into the following linear programming problem[16]:

$$\max_{m \in \mathcal{M}} \frac{1}{n/2} \sum_{i \in \mathbb{S}_{low}} \sum_{j \in \mathcal{S}_{high}} f(y_{i0}, y_{j0}, |n_{i0} - n_{j0}|) m(i,j) \tag{11}$$

such that $m : \mathbb{S}_{low} \times \mathbb{S}_{high} \to \mathbb{R}_+$ is the fraction of matches between type $i \in \mathcal{S}_{low}$ and $j \in \mathcal{S}_{high}$. The function m is to be chosen is restricted to be within a set $\mathcal{M}$ defined by the following constraints:

$$m(i,j) \geq 0, \forall i \in \mathcal{S}_{low}, \forall j \in \mathbb{S}_{high} \tag{12}$$

$$\sum_{i \in \mathcal{S}_{low}} m(i,j) = 1, \forall j \in \mathbb{S}_{high} \tag{13}$$

$$\sum_{j \in \mathcal{S}_{high}} m(i,j) = 1, \forall i \in \mathbb{S}_{low} \tag{14}$$

In particular, the two accounting constraints (13), (14) govern that the number of matches must match the mass of agents of each type. Since each type in a deskmate matching problem is a particular seat occupied by exactly one student, the total mass is 1 for each type. Noting that our problem is discrete, m can be regarded as a doubly stochastic matrix of dimension $(n/2) \times (n/2)$.[17] Even though we consider the relaxed problem, the optimal assignments remain pure.

We first run a set of peer regressions with full three-way interactions among own baseline score, peer baseline scores, and non-cognitive peer difference.

[16] In the language of optimal transport, our original problem is the Monge problem and our relaxed problem is a Kantorovich problem. See Villani (2008) and Galichon (2016).

[17] Linear programs of this size can be solved almost instantaneously by using non-commercial solvers, so that computation is a non-issue.

We do not discretize the own baseline score like we did previously in Section 4, which we did only for presentation purposes. We evaluate the academic achievement given four kinds of peer assignments:

1. (Full Interaction) Optimization based on the full interaction model.
2. (Univariate) An assignment based on optimizing a mis-specified model with only the two-way interaction between own and peer baseline scores; non-cognitive peer difference is ignored. This is the model considered in Carrell et al. (2013) and most academic peer effect papers. This univariate model is best-fitted to the same data.
3. (Worst Assignment) The value obtained by the worst assignment (minimizing the objective function).
4. (Random Assignment) The achievement averaged over 1000 random assignments.

We then compute the value of the true objective function (11) given the assignments.

We use random assignments as a benchmark and compare the other optimization results against it. We subtract the optimization results from that obtained from the randomized assignment for each class; then, we compute the first quartile, the median, and the third quartile of this difference. To be precise, let $a \in \{1, 2, 3\}$ where $1, 2, 3$ label the optimal assignment, univariate, worst assignment respectively; let c denote a particular class; let $\bar{y}_c^a$ denote the computed average score of low-achieving students in class c under assignment a. let $\bar{y}_c^{random}$ denote the average of average scores under 1000 random assignments. Our exercise is to evaluate the quantiles of $\bar{y}_c^a - \bar{y}_c^{random}$ for $a \in \{1, 2, 3\}$.

We repeat this procedure separately for Chinese mid-term and final, and the math equivalents. The results in Table 4 show that the univariate model is almost indistinguishable from the random benchmark, such that the first quartiles, the medians, and the third quartiles are all very close to zero.

Since the dependent variables are all log scores, the differences are all interpreted as percentages; for instance, the median increase in Chinese midterm grades obtained by optimizing according to our extended peer effect model is

4%. There are variation of this increase among classes, although all classes shows an improvement.

Another consistent pattern is that the benchmark is relatively close to the worst assignment than to the optimal assignment, thereby indicating that without significant care in optimizing the peer assignment, the optimized result cannot be easily obtained by chance.

Table 4: Optimized Achievement Scores

		Full Interaction	Univariate	Worst Assignment
Chinese Midterm	1st Quartile	0.028	0.006	-0.052
	Median	0.04	0.01	-0.035
	3rd Quartile	0.062	0.015	-0.026
Chinese Final	1st Quartile	0.028	-0.003	-0.04
	Median	0.041	0.005	-0.028
	3rd Quartile	0.064	0.013	-0.013
Math Midterm	1st Quartile	0.013	0.005	-0.048
	Median	0.019	0.008	-0.029
	3rd Quartile	0.024	0.011	-0.022
Math Final	1st Quartile	0.016	0.002	-0.038
	Median	0.025	0.005	-0.024
	3rd Quartile	0.030	0.007	-0.009

Notes: All results are relative to the results obtained by random assignment.

In terms of magnitudes, peer assignment optimization is more effective in improving Chinese (about 4%) than math (about 2%), which is expected given that math heavily depends on cognitive ability. These figures are modest yet expected since this is only a one semester experiment.

One way to interpret our objective function is to think of it as a trade off between non-linearity (the interaction between $y_{i0}, y_{q(i)0}$) and non-cognitive peer difference $|n_{i0} - n_{q(i)0}|$. In Table 5, we report the average non-cognitive difference under the various assignments. Indeed, we find that for Chinese midterm, Chinese final and Math final, the optimal assignments have an average non-cognitive peer difference of around $0.8 - 0.86$, which is about 14-20% less than the average non-cognitive peer difference found in the univariate or random assignments. In contrast, the worst assignment magnifies the non-cognitive peer difference in the class.

Math midterm is an exception. The non-linearity between the baseline scores is particularly strong in this case, so that the optimization algorithm tends to match students with a large non-cognitive peer difference in pursue for gains from non-linearity. However, Math Final does not exhibit the same phenomenon. One may interpret this result as one that emphasizes the long-run importance of non-cognitive peer difference.

Table 5: Non-Cognitive Peer Difference (Medians)

	Optimal Assignment	Univariate	Worst Assignment
Chinese Midterm	0.787	1.049	1.035
Chinese Final	0.806	1.028	1.052
Math Midterm	1.270	1.030	1.060
Math Final	0.865	0.977	0.991

Notes: All results are relative to the results obtained by random assignment.

In Table 6, we check whether the optimal assignment tends to assign female-female pairs. We compare these figures to the corresponding ones under random assignment and report their ratios. Again, we report the figures with respect to the four outcomes, and all figures are very close to 1 because, as Table 7 indicates, the baseline non-cognitive peer difference has almost zero systematic difference across gender pairs.[18]

Table 6: Number of Same-Sex Pairs (Medians)

	Optimal Assignment	Univariate	Worst Assignment
Chinese Midterm	1.017	1.038	0.981
Chinese Final	0.958	0.974	1.013
Math Midterm	0.977	1.046	0.984
Math Final	0.944	1.063	1.019

Notes: All results are relative to the results obtained by random assignment.

We then perform a hypothetical exercise where we replace $|n_{i0} - n_{q(i)0}|$ with $|n_{i1} - n_{q(i)1}|$, and examine if the optimal assignment significantly changes. If there is no assimilation such that non-cognitive peer difference is stable over time, then the optimal assignment should be stable as well. However, we find that apart from a very small number of exceptions, the optimal assignment $q(i)$ changes for all i (the percentage is $\geq 95\%$ for all classes). This hypothetical

[18] We also have tried to characterize the optimal assignment according to age, whose results are similar.

assignment demonstrates the need for readjustments instead of considering a one-shot optimization. Note that a class teacher, who monitors his class daily, would possess real-time information regarding non-cognitive peer difference. Athough he or she may not be able to accurately measure non-cognitive peer difference, he or she possesses significant advantage in having high-frequency measurements of non-cognitive peer difference that a peer effect researcher (the authors) do not have. Therefore, it would not be surprising if the class teachers perform better than our one-shot optimization attempt, which finds a modest gain of about 4%.

We compute the number female-female pairs for our hypothetical assignment and compare the figure against that of the optimal assignment. We do not find any systematic results to report. Males account for the majority in the low-achieving groups (62.0% on average), yet comprises the minority in the high-achieving group (45.6% on average) which agrees with the reverse gender gap literature. Consequently, the optimal assignment, that restricts the pairing to be between high- and low-achieving students, cannot form many female-female pairs.

Although our focus is the peer assignment betwen high- and low-achieving students, the same methodology can be applied to other kinds of assignments, for instance that between two randomized groups. The welfare weights may change so that the academic performance of high-achieving students are also considered. Since our purpose is to demonstrate the general principles, we do not elaborate on these possible variations.

6 Assimilation

We report the temporal change in non-cognitive peer difference $d_{i1} - d_{i0}$. Figure 2 plots its distribution. In the plot, 51% of the observations assimilate, i.e. $d_{i1} - d_{i0} \leq 0$, while the other 49% differentiate, i.e. $d_{i1} - d_{i0} > 0$. The distribution of $d_{i1} - d_{i0}$ is unimodal, with a mean of -0.03, which is statistically significant at the one percent significance level.

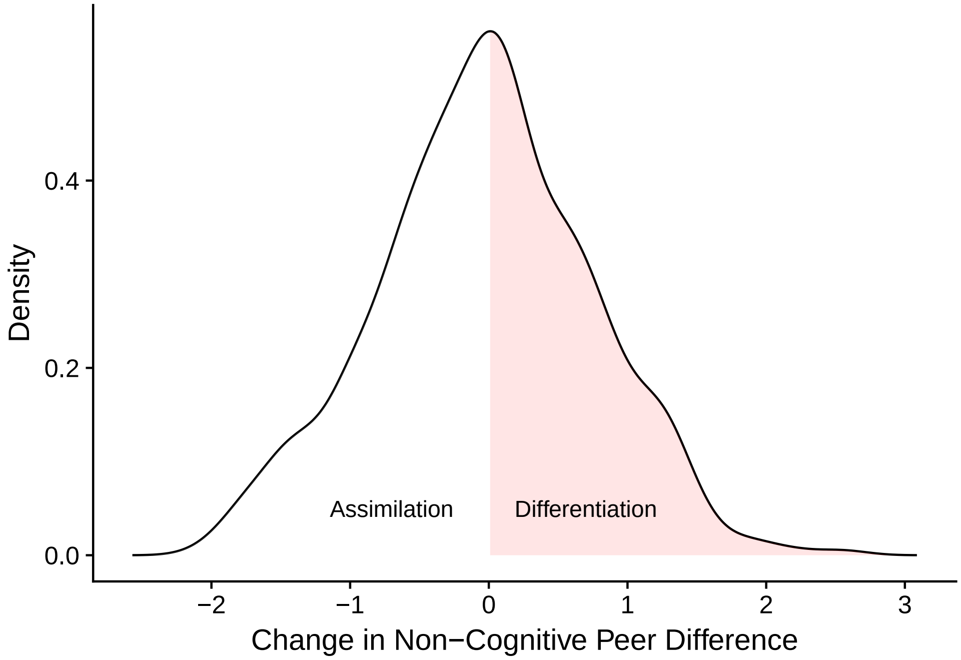


Figure 2: Distribution of Change in Non-Cognitive Peer Difference (Round 2 - Round 1)

6.1 By Gender

We find that the success of assimilation highly depends on own and peer gender. Table 7 reports how gender explains the differences of the non-cognitive skill between deskmates. Specifically, we run an OLS regression of d_{it} for $t = 0, 1$ and $d_{i1} - d_{i0}$ on possible gender pairs, i.e. male-female, male-male and female-female, with male-male as the omitted group.

At the baseline, there is almost zero level difference in d_{i0} across gender pairs. After a semester, we find that $d_{i1} - d_{i0}$ (Column 3) is heterogeneous across gender pairs. Specifically, female-female pairs has a significantly negative coefficient of -0.1747, which is about 6 times relative to the mean assimilation of 0.03. While this figure is only $1/5$ of the standard deviation of assimilation $d_{i1} - d_{i0}$, it is still significant. We do not find similar effects for the other gender pairs. In other words, the degrees of assimilation of mixed gender pairs and male-male pairs are similar.

These result relate to Lu and Anderson (2015), who find that female-female pairs improve academic achievement relative to other possible gender pairs. Together with our result that a smaller non-cognitive peer difference interacts with the peer baseline score effect, the significantly better assimilation of female-female pairs can potentially explain their results. Lavy and Schlosser (2011) also show that adding female students to a class can improve the overall academic performance. As the authors point out, a higher proportion of female peers "lowers the level of classroom disruption and violence, improves inter-student relationships." While their results are global and our gender assimilation result is local (between deskmates), our findings are consistent with each other.

Table 7: Explaining Non-Cognitive Peer Differences (By Gender)

	Baseline	After the First Semester	Change
	(1)	(2)	(3)
Male-Female	−0.0112	−0.0911*	−0.0799
	(0.0452)	(0.0510)	(0.0672)
Female-Male	−0.0038	−0.0995*	−0.0957
	(0.0450)	(0.0508)	(0.0669)
Female-Female	−0.0164	−0.1911***	−0.1747**
	(0.0555)	(0.0626)	(0.0825)
class FE	Yes	Yes	Yes
location FE	Yes	Yes	Yes
sd of Dep. Variable	0.5706	0.6449	0.7736
Observations	892	892	892
R^2	0.4745	0.4773	0.3695
Adjusted R^2	0.4125	0.4157	0.2952
Notes:	***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.		

6.2 By Grade

Similarly, we examine the non-cognitive peer difference by grade. The omitted group in Table 8 is Grade 3. We do not introduce class dummies because they would absorb all relevant variation.

At the baseline, the average non-cognitive peer differences of Grade 4 and Grade 5 are smaller than that of Grade 3, as indicated by their negative coefficients. This result suggests that initially the higher grades are more homogenous than the lower grades.

In terms of change, the positive Grades 4 and 5 coefficients suggest that the non-cognitive peer differences in higher grades are more difficult to be reduced relative to those of Grade 3. Similar to our interpretation for gender, the magnitudes of the coefficients (greater than 0.2 for both Grade 4 and Grade 5) are large with respect to the baseline non-cognitive peer difference with unit variance.

Our result also agrees with the skill formation literature, which shows that non-cognitive skill is more malleable at younger ages (and hence lower grades). While the findings in the skill formation literature mostly concern school and family inputs, here we show that the same conclusion holds with respect to peer inputs. Specifically, we show this result directly by measuring assimilation between peers. In the skill formation literature, while peer inputs are often mentioned, the data set used is a representative sample whose observations are i.i.d.. The peer inputs are measured indirectly by interviewing the respondent rather than his or her peer.

Table 8: Explaining Non-Cognitive Peer Differences (by Grade)

	Baseline	After the First Semester	Change
	(1)	(2)	(3)
Grade 4	-0.080^{*}	0.134^{***}	0.214^{***}
	(0.043)	(0.050)	(0.058)
Grade 5	-0.180^{***}	0.108^{**}	0.288^{***}
	(0.043)	(0.049)	(0.058)
location FE	Yes	Yes	Yes
sd of Dep. Variable	0.5706	0.6449	0.7736
Observations	892	892	892
R^2	0.095	0.088	0.109
Adjusted R^2	0.015	0.006	0.029

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

6.3 By Age Difference

We also examine within each class (and hence grades) how age difference affect the evolution of non-cognitive peer difference. The regression results are shown in Table 9. The first column shows that at the baseline, due to peer randomization, a larger age difference does not statistically predict a larger peer non-cognitive difference. However, after a semester, one extra year in peer age difference predicts a 6% increase in non-cognitive peer difference, suggesting that pairs with a larger age difference face more difficulties in assimilation.

Comparing with the gender results presented in Table 7, the effect of age difference on assimilation is not as large, although still notable and strongly significant. In Chinese rural schools, due to a number of reasons (mostly due to deferrals, but also include gifted children who are promoted to a higher class early), the age variation within classes is rather large. As reported in the summary statistics in Table 1, some classes have a within-class standard deviation of birth year exceeding 1 year, thereby implying that some pairs can have a large age difference relative to other pairs.

Table 9: Explaining Non-Cognitive Peer Differences (by Age)

	Baseline	After the First Semester	Change
	(1)	(2)	(3)
Age Difference	0.021*	0.059***	0.038**
	(0.012)	(0.014)	(0.018)
class FE	Yes	Yes	Yes
location FE	Yes	Yes	Yes
sd of Dep. Variable	0.5706	0.6449	0.7736
Observations	892	892	892
R^2	0.478	0.483	0.371
Adjusted R^2	0.418	0.423	0.299

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

6.4 Assimilation and Friendship

As we constructed, non-cognitive peer difference is a psychological factor score. This measure could be better understood if it relates observable consequences to the social network. Specifically, we check if differentation actually causes friendships to break. As another important concern, the non-cognitive questionnnaire items are self-reported. Given that the students are young in age, one may also wonder whether the students fully understood the questionnaire items, and whether they would seriously think about the best answer before answering. To address these concerns, in our sample, we collect information on friendships. Specifically, we ask each student to provide the names of three male friends and three female friends, six persons in total. Using this information, we check whether the deskmates are among the self-reported friends in each round, and check if this reporting relates to assimilation.

Initially, about 30% of the deskmates are named as one of the friends; after a semester (in Round 2, only 22% of the Round 1 deskmates are named as friends. These statistics altogether suggest that a student is likely to causally write down the name of his or her randomly assigned deskmate as one of his/her friends, given that there are six slots in total. In this case, one interesting issue is then what causes a student to stop writing his or her Round 1 deskmates again as a friend in Round 2, given that he/she did so in Round 1. If this is the case, we expect this behavior is a strong indicator of failed assimilation.

As Table 10 shows, non-cognitive peer difference does not predict whether a student would report his or her deskmate as a friend in Round 1. In Round 2, we define a dummy "assimilation" which takes a value of 1 if the change in non-cognitive peer difference is negative. The interaction between assimilation and reporting the Round 1 deskmate as a friend in Round 1 has a positive effect. Quantitatively, this result implies that the probability of reporting a Round 1 deskmate as a friend in Round 2, conditioned on that he did so in Round 1, is 17% lower if the assimilation fails (differentiation occurs).

As a caveat, this result is not a causal statement because the assimilation and change in friendship are concurrent. Therefore, we do not claim that being friends (stop being friends) is a result of assimilation (differentiation), or if the reverse causality channel is true. Instead, we prefer take this result to support our view that non-cognitive peer difference proxies whether peers have a close relationship to each other.

We do not have a direct measure of peer conflict. Although unfortunate, including this item in the questionnaire could raise ethical issues.

Table 10: Friendship and Deskmates: OLS Regressions

	Deskmate1 is Friend in Round 1	Deskmate1 is Friend in Round 2
	(1)	(2)
Non-Cog Diff (Round1)	−0.004	−0.005
	(0.014)	(0.012)
Assimilation		0.014
		(0.027)
Deskmate1 is Friend in Round 1		0.343^{***}
		(0.035)
Interaction (2)*(3)		0.169^{***}
		(0.053)
Constant	0.241^{***}	0.088^{***}
	(0.014)	(0.019)
Observations	892	892
R^2	0.000	0.124
Adjusted R^2	−0.001	0.120

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.
Non-cognitive peer difference in both rounds are standardized.

7 Conclusions

We end the paper with several remarks. We aim to shed light on the peer effect puzzle raised by Carrell et al. (2013), who show that the optimization of peer assignment could possibly backfire even if the estimates are obtained from an experiment. As our results show, non-cognitive peer difference is of critical importance. Therefore, we need to extend the standard academic peer effect model before we can recommend on actual peer assignments.

We begin by observing that Chinese class teachers routinely readjust the seating arrangement in their classes. Although their approach is often trial-and-error in nature, it is basically a response to peer conflicts between deskmates. Therefore, we question whether the existing academic peer effect models could have ignored a pre-condition that the assigned peers have to get along with each other well before knowledge spillover can take place effectively. To get along with each other well, their non-cognitive peer difference should be small. We quantify this idea by measuring non-cognitive peer difference, and then we optimize peer assignment based on our richer peer effect model. Our results are in line with our expectations.

Our paper focuses on a setting with a fixed seats, which provides us the possibly cleanest environment to study peer interactions. Concerning external validity, we not that the fixed seat system, as the traditional way of organizing seats for primary school classrooms, is common in other East Asian regions such as Singapore, Hong Kong, India, South Korea, and Japan. Similar systems can also be found in other regions where limited education resources demands a compact way of organizing classrooms. A fixed seat system is often found in workplaces with a limited office space; in these workplaces where the assignment of employees into fixed seats could result in peer conflicts if done improperly. However because workers are specialized in particular roles, thereby giving rise to the need for each worker to communicate with specific peers, the discussion of optimal seating assignment in workplaces involves is a more complicated problem than that in classrooms.

As a related remark, the rectangular array arrangement could produce global effects. Specifically, this approach can be used as a means of inducing obedience and order. As education researchers suggest, a semi-circular seating layout can effecitvely encourage students to ask questions to their teacher. By contrast, a rectangular array seating layout can promote the teacher as a authoritative figure (Marx et al., 1999). Our peer effect regressions control for such global effects by using classroom fixed effects. We leave to future studies the problem of globally changing the classroom seating layout.

References

ABRAMITZKY, R., L. P. BOUSTAN, AND K. ERIKSSON (2016): "Cultural assimilation during the age of mass migration," Tech. rep., National Bureau of Economic Research.

AIZER, A. (2008): "Peer effects and human capital accumulation: The externalities of ADD," Tech. rep., National Bureau of Economic Research.

AMMERMUELLER, A. AND J.-S. PISCHKE (2009): "Peer effects in European primary schools: Evidence from the progress in international reading literacy study," *Journal of Labor Economics*, 27, 315–348.

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

ANGRIST, J. D. AND V. LAVY (1999): "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *Quarterly Journal of Economics*, 114, 533–575.

BLACK, D. A. AND J. A. SMITH (2006): "Estimating the returns to college quality with multiple proxies for quality," *Journal of Labor Economics*, 24, 701–728.

BOOIJ, A. S., E. LEUVEN, AND H. OOSTERBEEK (2017): "Ability peer effects in university: Evidence from a randomized experiment," *The Review of Economic Studies*, 84, 547–578.

BORJAS, G. J. (1985): "Assimilation, changes in cohort quality, and the earnings of immigrants," *Journal of labor Economics*, 3, 463–489.

——— (1995): "Assimilation and changes in cohort quality revisited: what happened to immigrant earnings in the 1980s?" *Journal of Labor Economics*, 13, 201–245.

——— (2015): "The slowdown in the economic assimilation of immigrants: Aging and cohort effects revisited again," *Journal of Human Capital*, 9, 483–517.

CAMERON, A. C. AND D. L. MILLER (2015): "A practitioner's guide to cluster-robust inference," *Journal of Human Resources*, 50, 317–372.

CARRELL, S. E., R. L. FULLERTON, AND J. E. WEST (2009): "Does your cohort matter? Measuring peer effects in college achievement," *Journal of Labor Economics*, 27, 439–464.

CARRELL, S. E. AND M. L. HOEKSTRA (2010): "Externalities in the classroom: How children exposed to domestic violence affect everyone's kids," *American Economic Journal: Applied Economics*, 2, 211–28.

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, 855–882.

CHETTY, R., J. N. FRIEDMAN, N. HILGER, E. SAEZ, D. W. SCHANZENBACH, AND D. YAGAN (2011): "How does your kindergarten classroom affect your earnings? Evidence from Project STAR," *The Quarterly Journal of Economics*, 126, 1593–1660.

CHISWICK, B. R. (1978): "The effect of Americanization on the earnings of foreign-born men," *Journal of political Economy*, 86, 897–921.

CORNELISSEN, T., C. DUSTMANN, AND U. SCHÖNBERG (2017): "Peer effects in the workplace," *American Economic Review*, 107, 425–56.

COSTA, P. T. AND R. R. MACCRAE (1992): *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI): Professional manual*, Psychological Assessment Resources, Incorporated.

COSTA JR, P. T. AND R. R. MCCRAE (1992): "Four ways five factors are basic," *Personality and Individual Differences*, 13, 653–665.

CUNHA, F. AND J. HECKMAN (2007): "The technology of skill formation," *American Economic Review*, 97, 31–47.

CUNHA, F. AND J. J. HECKMAN (2008): "Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation," *Journal of Human Resources*, 43, 738–782.

CUNHA, F., J. J. HECKMAN, AND S. M. SCHENNACH (2010): "Estimating the technology of cognitive and noncognitive skill formation," *Econometrica*, 78, 883–931.

DING, W. AND S. F. LEHRER (2007): "Do peers affect student achievement in China's secondary schools?" *The Review of Economics and Statistics*, 89, 300–312.

DUFLO, E., P. DUPAS, AND M. KREMER (2011): "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya," *American Economic Review*, 101, 1739–74.

EPPLE, D., E. NEWLON, AND R. ROMANO (2002): "Ability tracking, school competition, and the distribution of educational benefits," *Journal of Public Economics*, 83, 1–48.

FALK, A. AND A. ICHINO (2006): "Clean evidence on peer effects," *Journal of labor economics*, 24, 39–57.

FELD, J. AND U. ZÖLITZ (2017): "Understanding peer effects: on the nature, estimation, and channels of peer effects," *Journal of Labor Economics*, 35, 387–428.

FIGLIO, D. N. (2007): "Boys named Sue: Disruptive children and their peers," *Education Finance and Policy*, 2, 376–394.

FIGLIO, D. N. AND M. E. PAGE (2002): "School choice and the distributional effects of ability tracking: does separation increase inequality?" *Journal of Urban Economics*, 51, 497–514.

FORTIN, N. M., P. OREOPOULOS, AND S. PHIPPS (2015): "Leaving boys behind gender disparities in high academic achievement," *Journal of Human Resources*, 50, 549–579.

FREDRIKSSON, P., B. ÖCKERT, AND H. OOSTERBEEK (2012): "Long-term effects of class size," *Quarterly Journal of Economics*, 128, 249–285.

FU, C. AND N. MEHTA (2014): "Ability tracking, school and parental effort, and student achievement: A structural model and estimation," Tech. rep., Working paper, accepted by Journal of Labor Economics.

GALICHON, A. (2016): *Optimal transport methods in economics*, Princeton University Press.

GOLDIN, C., L. F. KATZ, AND I. KUZIEMKO (2006): "The homecoming of American college women: The reversal of the college gender gap," *Journal of Economic perspectives*, 20, 133–156.

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, 34–68.

HANUSHEK, E. A. (1998): "Improving Student Achievement: Is Reducing Class Size the Answer?" Tech. rep., Policy Briefing, Progressive Policy Institute, Washington, DC, June.

——— (2008): "Education production functions," *The New Palgrave Dictionary of Economics. Basingstoke: Palgrave Macmillan*.

HANUSHEK, E. A., J. F. KAIN, J. M. MARKMAN, AND S. G. RIVKIN (2003): "Does peer ability affect student achievement?" *Journal of Applied Econometrics*, 18, 527–544.

HARRIS, J. R. (1995): "Where is the child's environment? A group socialization theory of development." *Psychological review*, 102, 458.

HECKMAN, J. J., J. STIXRUD, AND S. URZUA (2006): "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," *Journal of Labor economics*, 24, 411–482.

HENDERSON, J. V., J. QUIGLEY, AND E. LIM (2009): "Urbanization in China: Policy issues and options," Tech. rep., Unpublished manuscript, Brown University.

HONG, S. C. AND J. LEE (2017): "Who is sitting next to you? Peer effects inside the classroom," *Quantitative Economics*, 8, 239–275.

HOXBY, C. (2000): “Peer Effects in the classroom: Learning from Gender and Race variation,” Tech. rep., National Bureau of Economic Research.

HOXBY, C. M. AND G. WEINGARTH (2005): “Taking race out of the equation: School reassignment and the structure of peer effects,” Tech. rep., Working paper.

JIN, L. AND M. CORTAZZI (1998): “Dimensions of dialogue: Large classes in China,” *International Journal of Educational Research*, 29, 739–761.

KONYA, I. (2007): “Optimal immigration and cultural assimilation,” *Journal of Labor Economics*, 25, 367–391.

KRUEGER, A. B. (1999): “Experimental estimates of education production functions,” *Quarterly Journal of Economics*, 114, 497–532.

——— (2003): “Economic considerations and class size,” *The Economic Journal*, 113.

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

LAZEAR, E. P. (1999): “Culture and language,” *Journal of Political Economy*, 107, S95–S126.

——— (2001): “Educational production,” *The Quarterly Journal of Economics*, 116, 777–803.

LU, F. AND M. L. ANDERSON (2015): “Peer effects in microenvironments: The benefits of homogeneous classroom groups,” *Journal of Labor Economics*, 33, 91–122.

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

MARX, A., U. FUHRER, AND T. HARTIG (1999): “Effects of classroom seating arrangements on children’s question-asking,” *Learning Environments Research*, 2, 249–263.

MAS, A. AND E. MORETTI (2009): "Peers at work," *American Economic Review*, 99, 112–45.

MENG, X. AND R. G. GREGORY (2005): "Intermarriage and the economic assimilation of immigrants," *Journal of Labor economics*, 23, 135–174.

MOULTON, B. R. (1986): "Random group effects and the precision of regression estimates," *Journal of Econometrics*, 32, 385–397.

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

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

——— (2014): "Experimental and quasi-experimental analysis of peer effects: two steps forward?" *Annual Review of Economics*, 6, 253–272.

TENG, M. AND H. LIANG (2012): "A Survey of Current Conditions of Large-Number Classes in the Primary and Secondary Schools in Dengzhou City, Henan Province," in *Annual Report on China's Education, Beijing, China*, ed. by D. Yang, Social Sciences Acadmic Proess (China), in Chinese.

TODD, P. E. AND K. I. WOLPIN (2003): "On the specification and estimation of the production function for cognitive achievement," *The Economic Journal*, 113.

URQUIOLA, M. (2006): "Identifying class size effects in developing countries: Evidence from rural Bolivia," *Review of Economics and Statistics*, 88, 171–177.

VILLANI, C. (2008): *Optimal transport: old and new*, vol. 338, Springer Science & Business Media.

A Measurement Errors and Standard Errors

A.1 Measurement Error

Measurement error is a well-known problem in psychometrics, especially given the fact that non-cognitive traits are typically more difficult to measure than cognitive ones. A common way to tackle this problem is to combine many measurements to form an index, in order for the measurement errors in each questionnaire item to cancel each other.

Formally, suppose that the non-cognitive skill n_{it} is latent, and is linked to a vector of (observable) measurements $(\hat{n}_{it1}, \hat{n}_{it2}, \ldots, \hat{n}_{itK})$ for individual i at time k, where $k = 1, 2, \ldots, K$ indexes the questionnaire items. The linkage is specified by:

$$\hat{n}_{itk} = \rho_{tk} n_{it} + \varepsilon_{itk} \tag{15}$$

where ρ_{tk} is the factor loading, and ε_{itk} is the measurement error. Given that the measurement errors are independent to the factor, $\mathbb{E}[\varepsilon_{itk} | n_{it}, t] = 0$ for any (t, k).

In this case, peer randomization at time 0 guarantees that:

$$(Y_{i0}, \varepsilon_{i01}, \varepsilon_{i02}, \ldots, \varepsilon_{i0K}, u_{i0}) \perp\!\!\!\perp (Y_{p(i)0}, \varepsilon_{p(i)01}, \varepsilon_{p(i)02}, \ldots, \varepsilon_{p(i)0K}, u_{p(i)0}) \tag{16}$$

Accordingly, we proxy d_{it} by

$$\hat{d}_{it} \equiv \sum_{k=1}^{K} \frac{|(\hat{n}_{itk} - \hat{n}_{p(i)tk})\delta_{tk}|}{K}$$

where $\boldsymbol{d}_t = (d_{t1}, d_{t2}, \ldots, d_{tK})'$ is a vector of weights. As such, we reduce the problem to one that has one single latent variable d_{it}, which is proxied by $\hat{d}_{it}$.

The literature on measurement errors discusses alternative methods other than simple averaging. The first is to optimize the weights in minimizing the attenuation bias. If one is willing to make the stronger asumption of independence between measurement errors across questionnaire items, as most factor models do, then

$$\varepsilon_{itk} \perp\!\!\!\perp \varepsilon_{itk'}, \varepsilon_{j(i,t)tk} \perp\!\!\!\perp \varepsilon_{j(i,t)t,k'} \forall k \neq k'$$

This assumption implies that the latent factor account for all cross-measurement correlations. Given this assumption, simple average tends to average out the measurement errors, although the weights may not be optimal because the measurement errors may not necessarily have the same variances.

The econometric problem of optimal weighting is discussed in Black and Smith (2006), who measured an unobserved college quality with multiple noisy proxies. Black and Smith (2006) estimated an optimized δ, found by minimizing the resulting expected mean deviation from the true latent factor. Nevertheless, even after optimizing the weights, measurement error still exists. Black and Smith (2006) finds that, after optimizing the weights, the factor score has statistically significant effects of college quality on various individual outcomes. In our case, we already find a very large effect without optimizing the weights. Being conservative, we choose not to optimize the weights and instead let the estimates to be attenuated.

Alternatively, the cross-measurement independence yields enough moments to achieve point identification of the factor loadings. This approach has also been employed in the skill formation literature. Heckman et al. (2006); Cunha et al. (2010) examined the joint identification of their model without calculating a factor score for each individual. Here we show the point identification of our peer effect model (for a linear case).

Conditional on $Y_{i0}, Y_{j(i,0)0}$, we compute the following partial variance-covariance matrix:

$$\begin{bmatrix} \tilde{var}(Y_{i1}) & \tilde{cov}(Y_{i1}, \hat{d}_{i01}) & \dots & \tilde{cov}(Y_{i1}, \hat{d}_{i0K}) \\ \tilde{cov}(\hat{d}_{i01}, Y_{i1}) & \tilde{var}(\hat{d}_{i01}) & \dots & \tilde{cov}(\hat{d}_{i01}, \hat{d}_{i0K}) \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{cov}(\hat{d}_{i0K}, Y_{i1}) & \tilde{cov}(\hat{d}_{i0K}, \hat{d}_{i01}) & \dots & \tilde{var}(\hat{d}_{i0K}) \end{bmatrix}$$
$$= \begin{bmatrix} \beta_3^2 \tilde{var}(\hat{d}_{i0}) + \tilde{var}(u_{i0}) + \tilde{var}(u_{j(i,0)0}) & \beta_3 \rho_{01} \tilde{var}(\hat{d}_{i0}) & \dots & \beta_3 \rho_{0K} \tilde{var}(\hat{d}_{i0}) \\ \beta_3 \rho_{01} \tilde{var}(\hat{d}_{i0}) & \rho_{01}^2 \tilde{var}(d_{i0}) + 2\tilde{var}(\varepsilon_{i01}) & \dots & \rho_{01}\rho_{0K} \tilde{var}(\hat{d}_{i0}) \\ \vdots & \vdots & \ddots & \vdots \\ \beta_3 \rho_{0K} \tilde{var}(d_{i0}) & \rho_{0K}\rho_{01} \tilde{var}(d_{i0}) & \dots & \rho_{0K}^2 \tilde{var}(d_{i0}) \end{bmatrix} \quad (17)$$

Standardize $\rho_{01} = 1$, such that the latent non-cognitive skill N_{it} shares the same unit as the first measurement $\hat{d}_{i01}$. Then the remaining factor loadings

$\rho_{02}, \rho_{03}, \ldots, \rho_{0K}$ are identified. For example,

$$\rho_{02} = \frac{c\tilde{o}v(Y_{i1}, \hat{d}_{i02})}{c\tilde{o}v(Y_{i1}, \hat{d}_{i01})} \tag{18}$$

Given that $c\tilde{o}v(\hat{d}_{i01}, \hat{d}_{i02}) = \rho_{02} var(d_{i0})$, $c\tilde{o}v(Y_{i1}, \hat{d}_{i02}) = \beta_3 \rho_{02} v\tilde{a}r(d_{i0})$

$$\beta_3 = \frac{c\tilde{o}v(Y_{i1}, \hat{d}_{i02})}{c\tilde{o}v(\hat{d}_{i01}, \hat{d}_{i02})} \tag{19}$$

i.e. β_3 is identified as the instrumental variable estimator, by using the second measurement to instrument the first. After β_3 is identified, $\beta_0, \beta_1, \beta_2$ can be identified by regressing $Y_{i1} - \beta_3 \hat{d}_{i0}$ on $Y_{i0}, Y_{j(i,0)0}$. Given that other measurements are available, the system is over-identified.

As a result of peer randomization, $N_{i0} \perp\!\!\!\perp N_{j(i,0)0}$ and so are their respective measurements. Finally, with the factor loadings identified, a standard application of the Kotlarski's lemma identifies the distribution of d_{i0}.

As a related remark, measurement errors may directly affect the optimization since the policy-maker may not observe d_{i0} perfectly. In this case, the optimization problem is one that involves Bayesian updating. Let the information set at time 0 denoted by Ω_0. Ω_0 would represent the prior that the policy-maker knows in advance, specifically the aggregate-level patterns of non-cognitive peer differences by observable characteristics such as age and gender. The corresponding Bayesian optimization problem is presented as follows:

$$\max_{q \in \mathbb{Q}} \mathbb{E}[\frac{1}{n/2} \sum_{i \in \mathbb{S}_{low}} [f(y_{i0}, y_{q(i)0}, |n_{i0} - n_{q(i)0}|)] | \Omega_0]$$

Upon receiving further information, the information set Ω_0 would change, thereby providing a reason of readjusting the optimal assignment en route.

A.2 Clustered Standard Errors and Wild Bootstrap

Unlike standard representative samples, peer data sets are inherently non i.i.d. because peers (classmates) interact with each other. Since a peer effect model asserts that both own and peer observable inputs would jointly determine outcomes, the same needs to be considered for unobservable inputs as well. This consideration would imply that the composite error terms $u_{i0} + \lambda_p u_{p(i)0}$ are correlated across individuals within the same class. Specifically, given that the deskmate of one's deskmate is oneself, a mechanical correlation between the composite error term across individuals exists even if $u_{i0} \perp\!\!\!\perp u_{p(i)0}$.

The violation of the i.i.d. assumption implies that the OLS standard error is incorrect, often understating the truth (Moulton, 1986). A well-known solution to this problem is to cluster the standard error by class. One problem in practice is that the cluster-robust standard errors (CRSE) are derived by assuming asymptotics with respect to the number of clusters (rather than individuals), while in field experimental studies the number of clusters is often small. In Lu and Anderson (2015), only 12 classes are available, while in our study we have 21.[19] For a small-cluster correction, we follow Cameron and Miller (2015) and use cluster wild bootstrap standard error in our regressions, which is a way to produce bootstrap re-samples by swapping residuals within clusters.

Nevertheless, we find that the cluster wild bootstrap standard errors are rather close to CRSE, such that the qualitative conclusions of testing tne null hypothesis of zero peer effects are unaffected. Meanwhile the OLS standard errors are about half of the cluster-bootstrap standard errors, thereby implying that our corrections are necessary.

B Randomization Check

As discussed above, randomization guarantees the independence between own and peer variables, whether observable or not. For observables, this condition is testable. Given that the randomization in this study is done by the schools following our instructions, the purpose of this randomization check is to confirm whether the teachers have gone through the randomization procedure as instructed.

Following Sacerdote (2001), we check whether the own and deskmate baseline non-cognitive skills are uncorrelated given the baseline randomization, conditional on location (row × column) and class fixed effect, since the random-

[19] In principle, the cluster unit should be the randomization blocks. However, the blocks include a very limited number of observations, thereby rendering the estimation of cluster-specific variances unreliable.

ization is conditional on these variables. The corresponding regression results are organized in Table 11. All Big Five proxies have passed the randomization check, with the p-values all greater than 0.05. We also present in the same table the results for baseline test scores and birth year; these variables also pass the randomization check.

As the only exception, the male dummy fails the randomization check, with a coefficient of -0.33 and a standard deviation of 0.03. This result implies that the deskmate is more likely of the opposite gender than of the same gender.

Table 11: Randomization Check

	deskmate coeff	deskmate (se)	deskmate (p-value)
Openness	-0.05	0.04	0.18
Conscientiousness	-0.07	0.03	0.05
Extraversion	0.04	0.04	0.23
Agreeableness	-0.05	0.04	0.15
Neuroticism	-0.03	0.04	0.42
Birth Year	-0.06	0.04	0.07
Male	-0.33	0.03	0.00
Chinese Language (Baseline)	0.03	0.04	0.46
Math (Baseline)	0.03	0.04	0.33

This abnormality warrants a discussion. As one possibility, it could be due to a "small urn problem" as discussed in Guryan et al. (2009) — Specifically, given own gender, the remaining pool of students in the same block are more likely to be of the opposite gender because the sampling is without replacement. The resulting regression coefficient with respect to baseline gender would be negative. In our case, each block is relatively small, mostly with less than 20 students. Therefore, this problem is even more acute than that in Guryan et al. (2009), which studies golfer teams of bigger size. Guryan et al. (2009) proposed a fix to this small urn problem, that is, to add an additional control of the proportion of males within a block (i.e., the eligible peers) excluding own. As Guryan et al. (2009) pointed out, this solution requires significant block size variation to work. In our case, we find insufficient variation in block size within a class. Therefore, we do not attempt to use this solution.

The second possibility is that the randomization is, in fact, conditional on gender given the strong preference of teachers for mixed gender pairs. In any case, we wish to identify peer effects other than the gender effects found in Lu and Anderson (2015). Therefore, for all peer effect regressions in Section 4, we control for gender pairs.

C Placebo Test of Local Peer Effects

In Round 2 which takes place after the first semester, the class teachers have become more knowledgeable about their classes. Specifically, these class teachers typically identify certain pairs of deskmates who are in serious conflict with each other. Consequently, these class teachers typically ask to assign seats according to their own initiative. Given this situation, we ask the class teacher to globally reassign the seats again, although we do not enforce true randomization as we did in Round 1.

We find that 95% of the students have changed seats, i.e. their (x, y) coordinates become different in Round 2. Given that each randomization block is small in size, being assigned exactly the same seat in both semesters remains a statistical possibility. Regarding the x-coordinates, the randomizations are conditional on blocks that comprises two rows each. Under the assumption that the ordering of heights of students remains constant within a semester, the change in the x coordinate should be within $\{-1, 0, 1\}$ for all students. We find that this conclusion is true for 87.86% of the sample. Moreover, we do find that some students are re-assigned more than 3 rows away from his Round 1 position.

With the peer assignment no longer random, we cannot use Round 2 data to repeat our estimation of peer effects. Instead, since the deskmates for most students change in Round 2 we use the opportunity that the deskmates, we use this opportunity to conduct a placebo test. If peer effects are local, then the Round 1 deskmate variables should not correlate to the Round 2 outcomes.

We cannot reject the null that the Round 1 peers have zero effect. We add the Round 2 peers in the specification and the results are unaffected. These results are shown in Table 12. As expected, the estimated interaction effects are much smaller in magnitude, have mixed signs, and are all statistically insignificant.

Together with our previous finding that general neighbors do not matter, we argue that within-classroom peer effects are local in nature. This property is important for our policy recommendations. If within-classroom peer effects are global, then seat reassignments would be irrelevant.[20]

D Consistency of d_{i0} Across Traits

As a robustness check, we examine whether our conclusions critically depends on the definition of the non-cognitive peer difference. This is especially important given that our measure is a composite one that mixes the Big Five. In response, we compute a correlation matrix between the trait-specific $d_{i1} - d_{i0}$. The high correlations (about 0.8) in Table 13 indicate that inter-peer non-cognitive differences are consistent across the Big Five traits. If one pair has large peer difference relative to the average with respect to one Big Five trait, e.g. openness to experience, then a have large peer difference with respect to other traits (e.g. conscientiousness) is likely to be observed as well.

Table 14 presents an analogue correlation matrix for the second round. The correlations unambiguously increase by about 0.05, thereby suggesting that the non-cognitive peer differences become more consistent across factors over time. Given that the second round survey is done after the re-randomization of seats, our findings reject an alternative hypothesis that such high correlations is due to some deskmate pairs copying answers from each other.

Consequently, our peer effect estimates do not depend on whether a specific trait is used. Given that most economists pay special attention on conscientiousness, we report the same peer effect regressions using the conscientiousness peer differences instead in Table 15 and 16 respectively. The signs and magnitudes of the estimates are consistent to our main results.

As another robustness check, this appendix section provides the academic peer effect regression estimates if we replace the Big Five measurements with only conscientiousness. The results are similar to our main regressions.

[20] Given the independence between deskmates and more general neighbors, the Round 1 specifications can omit the general neighbor variables.

Table 12: Local Peer Effect Placebo Test (Round 2 Outcomes, Round 1 Peers)

	Chinese Midterm	Chinese Final	Math Midterm	Math Final
	(1)	(2)	(3)	(4)
Own Baseline	0.680***	0.701***	0.939***	0.982***
	(0.086)	(0.109)	(0.077)	(0.074)
Deskmate1 Baseline	−0.007	−0.002	−0.023	0.029
	(0.022)	(0.024)	(0.040)	(0.046)
Non-Cog Diff (Deskmate1)	0.047	−0.085	0.139	0.212
	(0.060)	(0.063)	(0.278)	(0.279)
Neighbor1 Baseline	−0.066	−0.059	0.039	−0.011
	(0.077)	(0.084)	(0.056)	(0.064)
Non-Cog Diff (Neighbor1)	0.083	−0.152	−0.054	0.212
	(0.130)	(0.106)	(0.160)	(0.200)
Interaction (Deskmate1)	−0.018	0.014	−0.030	−0.049
	(0.015)	(0.016)	(0.064)	(0.063)
Interaction (Neighbor1)	−0.015	0.038	0.007	−0.054
	(0.029)	(0.025)	(0.037)	(0.047)
Constant	1.400**	1.651**	−0.058	−0.222
	(0.630)	(0.703)	(0.533)	(0.588)
class FE	Yes	Yes	Yes	Yes
location2 FE	Yes	Yes	Yes	Yes
gender pair FEs	Yes	Yes	Yes	Yes
Observations	877	874	877	876
R^2	0.704	0.668	0.689	0.623
Adjusted R^2	0.660	0.618	0.642	0.567

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.
Standard Errors are obtained by wild bootstrap and clustered at class level.
All Test Scores are in logs.
Baseline Non-Cognitive peer differences are standardized.

Table 13: Correlation Matrix of Peer Differences in Factor-Specific Indexes (Baseline)

	O (Diff)	C (Diff)	E (Diff)	A (Diff)	N (Diff)
O (Diff)	1.00	0.82	0.84	0.81	0.80
C (Diff)	0.82	1.00	0.80	0.78	0.75
E (Diff)	0.84	0.80	1.00	0.81	0.80
A (Diff)	0.81	0.78	0.81	1.00	0.80
N (Diff)	0.80	0.75	0.80	0.80	1.00

Table 14: Correlation Matrix of Peer Differences in Factor-Specific Indexes (Second Round)

	O (Diff)	C (Diff)	E (Diff)	A (Diff)	N (Diff)
O (Diff)	1.00	0.86	0.89	0.86	0.85
C (Diff)	0.86	1.00	0.85	0.84	0.85
E (Diff)	0.89	0.85	1.00	0.86	0.87
A (Diff)	0.86	0.84	0.86	1.00	0.85
N (Diff)	0.85	0.85	0.87	0.85	1.00

Table 15: Peer Effect (Chinese, Conscientiousness)

	First Sem. Midterm			First Sem. Final		
	(1)	(2)	(3)	(4)	(5)	(6)
Own Baseline	0.774^{***}	0.765^{***}	0.777^{***}	0.592^{***}	0.552^{***}	0.824^{***}
	(0.047)	(0.043)	(0.072)	(0.058)	(0.058)	(0.102)
Deskmate1 Baseline	-0.029	-0.038	-0.021	0.016	-0.020	0.013
	(0.025)	(0.038)	(0.013)	(0.021)	(0.017)	(0.019)
Non-Cog Diff (Deskmate1)	-0.027	-0.055	-0.017	0.307^{***}	0.484^{***}	-0.001
	(0.103)	(0.133)	(0.036)	(0.062)	(0.140)	(0.035)
Interaction (Deskmate1)	0.003	0.003	0.004	-0.072^{***}	-0.117^{***}	0.001
	(0.026)	(0.037)	(0.008)	(0.015)	(0.038)	(0.008)
Interaction (Neighbor1)	1.110^{***}	1.114^{***}	1.039^{***}	1.682^{***}	2.026^{***}	0.673
	(0.190)	(0.235)	(0.336)	(0.199)	(0.253)	(0.486)
class FE	Yes	Yes	Yes	Yes	Yes	Yes
location FE	Yes	Yes	Yes	Yes	Yes	Yes
gender pair FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	891	450	455	888	448	454
R^2	0.632	0.642	0.492	0.689	0.717	0.544
Adjusted R^2	0.579	0.532	0.334	0.644	0.629	0.402

Notes:
***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.
Standard errors are obtained by wild bootstrap and clustered at class level.
All Test Scores are in logs.
(1),(4): Whole sample
(2),(5): Own baseline below the median
(3),(6): Own baseline above the median
Baseline non-cognitive peer differences are standardized.

Table 16: Peer Effect (Math,Conscientiousness)

	First Sem. Midterm			First Sem. Final		
	(1)	(2)	(3)	(4)	(5)	(6)
Own Baseline	0.840^{***}	0.817^{***}	1.212^{***}	0.839^{***}	0.874^{***}	0.981^{***}
	(0.058)	(0.070)	(0.101)	(0.079)	(0.085)	(0.076)
Deskmate1 Baseline	-0.023	-0.053	-0.037^{*}	0.057	0.051	-0.027^{**}
	(0.034)	(0.049)	(0.019)	(0.049)	(0.076)	(0.013)
Non-Cog Diff (Deskmate1)	0.055	0.266^{*}	-0.073	0.362^{*}	0.721^{**}	-0.009
	(0.091)	(0.160)	(0.071)	(0.208)	(0.362)	(0.038)
Neighbor1 Baseline	-0.008	-0.054	0.017	-0.081^{*}	-0.162^{**}	0.003
	(0.020)	(0.036)	(0.016)	(0.046)	(0.082)	(0.008)
Non-Cog Diff (Neighbor1)	0.556^{**}	0.674^{**}	-0.987^{**}	0.350	0.232	0.094
	(0.236)	(0.279)	(0.450)	(0.468)	(0.538)	(0.360)
class FE	Yes	Yes	Yes	Yes	Yes	Yes
location FE	Yes	Yes	Yes	Yes	Yes	Yes
gender pair FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	891	469	448	888	467	447
R^2	0.758	0.784	0.539	0.637	0.669	0.469
Adjusted R^2	0.724	0.720	0.392	0.585	0.571	0.300

Notes:

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.
Standard errors are obtained by wild bootstrap and clustered at class level.
All Test Scores are in logs.
(1),(4): Whole sample
(2),(5): Own baseline below the median
(3),(6): Own baseline above the median
Baseline non-cognitive peer differences are standardized.