MELBOURNE INSTITUTE
Applied Economic & Social Research

Working Paper Series

On the Mechanisms of Ability Peer Effects

Alexandra de Gendre
Nicolás Salamanca

Working Paper No. 19/20
October 2020

THE UNIVERSITY OF MELBOURNE
MELBOURNE INSTITUTE

On the Mechanisms of Ability Peer Effects*

Alexandra de Gendre
School of Economics, The University of Sydney
Institute of Labor Economics (IZA)

Nicolás Salamanca
Melbourne Institute: Applied Economic & Social Research,
The University of Melbourne
Institute of Labor Economics (IZA)

Melbourne Institute Working Paper No. 19/20
October 2020

*We thank David Figlio, Olivier Marie, and especially Jan Feld and Ulf Zölitz for their helpful comments. Boer Xia provided excellent research assistance for this project. We also thank the Survey Research Data Archive (SRDA) in Taiwan for providing us with data access, and Wan-wen Su for facilitating our analyses of the Taiwan Education Panel Survey. This research was supported by the Australian Research Council through the Centre of Excellence for Children and Families over the Life Course (project number CE140100027). The Centre is administered by the Institute for Social Science Research at The University of Queensland, with nodes at The University of Western Australia, The University of Melbourne and The University of Sydney. The views expressed herein are solely those of the authors.

Melbourne Institute: Applied Economic & Social Research
The University of Melbourne
Victoria 3010 Australia
***Telephone* +61 3 8344 2100**
***Fax* +61 3 8344 2111**
***Email* melb-inst@unimelb.edu.au**
***Website* melbourneinstitute.unimelb.edu.au**

Abstract

Studying with higher ability peers increases student performance, yet we have little idea why. We exploit random assignment of students to classrooms and find positive peer effects on test scores. With very rich data on seventeen potential mechanisms, we then estimate how peer effects on attitudes, parents, etc. could drive these results. Higher-achieving peers reduce student effort, increase student university aspirations, increase parental time investments, and have precise null effects elsewhere. None of these mechanisms, however, explain our peer effect on test scores. Our findings question the prevailing empirical approach to understanding the mechanisms underlying academic peer effects.

JEL classification: I23, I26, D13

Keywords: Random assignment; standardized test; mediation analysis; parental investment; school inputs

1 Introduction

Despite an immense literature in economics documenting the importance of peers for academic achievement, we still know little about their mechanisms. This remains an important limitation in our understanding of the theoretical underpinnings of peer effects, and limit our scope for using class assignment policies as a tool to improve student achievement and educational outcomes (Carrell, Sacerdote & West, 2013; Sacerdote, 2014; Ushchev & Zenou, 2020).

One key reason why it is difficult to make headway in understanding the mechanisms behind peer effects is that this requires large amounts of data from several sources. The prevailing view is that educational achievement is the output of an education production function which comprises several simultaneous inputs from students, parents and teachers (e.g. Cunha & Heckman 2007). If so, it is also natural to think that peers can affect educational achievement through any of these inputs. Yet datasets that collect information on all, or even many, of these inputs are rare. Because of this limitation, the prevailing approach in the empirical literature on academic peer effects is to focus on the effect of higher-achieving peers on the few available outcomes and discuss why these isolated responses may be mechanisms through which academic peer effects operate. Formal analyses of the share of peer effects we can explain are often futile with so few mechanisms to look at, leaving the key unanswered question: what drives academic peer effects?

To answer this question, we need three key elements *in the same environment*. First, we need to establish their existence; that there is a causal effect of being exposed to higher-achieving peers on students' own academic achievement. Second, we need to observe many candidate mechanisms that affect academic achievement. And third, we need to determine whether better peers improve academic achievement by shifting those factors. Many studies have gained access to one or two of these key elements, yet to date there is no study that provides empirical evidence all three of them jointly. We fill this gap.

In this paper, we first show the existence of academic peer effects, as others have done in different settings (for excellent reviews, see Sacerdote, 2011; 2014). We exploit a mandate to randomly assign student to classrooms within schools in our setting as a cornerstone in our identification strategy, and develop a method to identify and use only schools that adhere to this mandate in our analyses. We find that a one standard deviation (1 SD) increase in the average test scores of

classroom peers at baseline increases own test scores by 5.4 percent of a standard deviation two years later.

Using rich data on students, parents, teachers and schools, we then estimate the causal effect of higher-achieving peers on a large battery of student, parent, and teacher educational inputs, which are all potential mechanisms of academic peer effects. Having 1 SD higher-achieving academic peers decreases students' school effort, increases students' university aspirations, and increases their expected ability to go to university. Higher-achieving peers also increase parents' time investments. We do not find effects of higher-achieving peers on students' initiative in class, truancy, exam cheating, or academic self-efficacy. We also find no effects on parental investments in private tutoring, on strictness, emotional support or harsh parenting, or on parental aspirations for their child to go to university. Finally, we also find no effects on students' perceptions of their school environment or their teacher engagement. Some of the effect we do find complement existing evidence in the peer effect literature (Feld & Zölitz, 2017 on perceived quality of peer interactions; Bursztyn & Jensen, 2015 and Bursztyn, Egorov & Jensen, 2018 on social pressure and effort provision; Carrell, Hoekstra & Kuka, 2018 on long-run college enrolment). Yet most of our estimates explore unstudied mechanisms behind academic peer effects; in fact, no study before has been able to test as many candidate mechanisms as we do.

Combining our estimates of higher-achieving peers on score and on educational inputs, we then answer the question: How much of the academic peer effect can be explained by our measured mechanisms? To do this, we begin by estimating the returns of all our educational inputs on academic achievement using high-quality cumulative value-added models (Todd & Wolpin, 2007; Fiorini & Keane, 2014). Our estimates show large returns to many of our explored inputs. We then use these returns to map the effects of higher-achieving peers on educational inputs to academic achievement using mediation analyses (Gelbach, 2016). Our estimates show that our battery of educational inputs mediate a *negative* share of our academic peer effects—which means that the effects of higher-achieving peers on educational inputs make it harder, not easier, to explain academic peer effects. This negative mediation is largely driven by the combine negative effect of higher-achieving peers on student effort and its positive value-added returns. Our other inputs explored have a virtually null contribution to mediation. This is a surprising and important new

result that questions the prevailing empirical approach to understanding the mechanisms behind peer effects.

Finally, we perform an extensive set of sensitivity analyses for our results including: additional tests for conditional random assignment, alternative estimates with an exhaustive set of controls, calculations of the degree of correlated unobserved heterogeneity needed to explain away our findings, corrections for measurement error in student ability and for incomplete sampling of classrooms, inference corrections using randomization inference and multiple hypotheses testing adjustments, and an extensive exploration of heterogeneity in peer effects and their mediation.

Our paper makes several contributions to better understand the complex nature of academic peer effects. This is the first paper to provide a thorough test of the many possible mechanisms underlying academic peer effects, testing 17 of them covering all key agents in educational production. Most previous studies test some potential mechanisms for academic peer effects but never many at a time, and never in a formal mediation analysis (two other studies such as Gong, Lu & Song, 2019 and Zölitz & Feld, forthcoming, use this approach for exploring mechanisms behind peer gender effects). This is an important limitation since the many inputs in the education production function imply equally many mechanisms for peer effects to work through, and the only way to know how well we can explain peer effects is to jointly test all these potential mechanisms (see e.g. Bursztyn & Jensen, 2015). The fact that after our efforts we still do not know how academic peer effects work is a testament to their complexity. Our findings should spur future research to come up with new hypothesized and untested mechanisms, or alternative methods for exploring the currently-observed ones.

We also make two methodological contributions to the empirical literature on peer effects. First, we develop a well-defined algorithmic approach to conducting balancing tests and identifying non-compliers in quasi-experimental peer effect designs. This is particularly useful in settings with likely *partial compliance* to random assignment of students to classrooms and no fully reliable indicators of non-compliance. In such settings, researchers often try and account for systematic violations of random assignment by controlling for additional characteristics beyond balancing characteristics, which complicates the interpretation of peer effect estimates and weakens identification strategies. Our approach is a transparent alternative to improve the validity of quasi-

experimental research designs based on conditional random assignment without relying on conditioning pre-treatment covariates to account for failed randomization. Second, we provide a simple algorithm for randomization inference that observes the data structure of students within schools and within classrooms. Maintaining the data structure and, in particular, rigorously respecting assigned classroom sizes is crucial for correctly calculating permutation-based t-randomization p-values (Young, 2019) and for producing permutation-based tests of random assignment, which are commonly used in the empirical peer effects literature. We provide *Stata* code for these two procedures upon request.

2 Peer effects in education

Economists have been interested in peer effects for a long time, and have published over 100 articles in economic journals since 2009 on peer effects in education alone, 28 of them in Top Five journals.[1] One reason for the widespread interest in peer effects is that they could "be harnessed to cost-effectively improve public […] services" (BenYishay & Mobarak, 2018). In other words, the existence of peer effects implies a social multiplier effect. Inspired by this promise of peer effects, an immense empirical literature rose to provide evidence on their existence and size—notable in education but in other fields as well. After two decades of studies, the existence of peer effects in education is a well-established fact.

Peer effects are notoriously difficult to identify for two main reasons (Manski, 1993): self-selection into peer groups (i.e., that similar people sort into the same groups) and the reflection problem (i.e., that estimates capture both my effect on my peers and the effect of my peers on me). Self-selection introduces bias in peer effects estimates arising from omitted variables. Reflection ties together the effects of (endogenous) peer interactions with the effect of (exogenous) peer characteristics, complicating the interpretation of peer effect estimates.

[1] For brevity, we focus on studies of peer effects on academic achievement, but many other studies also document peer effects in e.g. college dropout (Stinebrickner and Stinebrickner, 2001), cheating in school (Carrell, Malmstrom, and West, 2008), job search (Marmaros, and Sacerdote, 2002), substance abuse (Argys and Rees, 2008; Kremer and Levy, 2008; Card and Giuliano, 2013), crime (Deming, 2011), technology adoption (Oster and Thornton, 2012), consumption (Moretti, 2011), financial decisions (Ahern, Duchin and Shumway, 2014; Bursztyn, Ederer, Ferman, Yuchtman, 2014) and beliefs (Boisjoly, Duncan, Kremer, Levy and Eccles, 2006).

Empirical studies typically solve the reflection problem by estimating the reduced-form effect of pre-assignment peer characteristics on student outcomes. Many studies have in addition convincingly solved the issue of self-selection by exploiting quasi-experimental assignment of students to peer groups. Two types of identification strategies have mainly been used to that end. The first strategy leverages (conditional) random assignment to peer groups within an institution. Examples include roommate assignment in college (Sacerdote, 2001; Stinebrickner & Stinebrickner 2001, 2006; Zimmerman, 2003; Foster, 2006; Brunello, De Paola & Scoppa, 2010; Griffith & Rask 2014; Jain & Kapoor, 2015; Garlick, 2018), classroom/section/dorm assignment within institutions (e.g., Lyle, 2007; Kang, 2007; Graham, 2008; Carrell, Fullerton and West, 2009; De Paola and Scoppa, 2010; Burke & Sass, 2013; Carrell, Sacerdote & West, 2013; Brady, Insler, & Rahmam, 2017; Feng & Li, 2016; Feld & Zölitz, 2017; Huntington-Klein & Rose, 2018; Garlick, 2018), and study group assignment within classroom (Lu & Anderson, 2014, Hong & Lee, 2017). The second identification strategy uses natural variation in cohort composition. Examples include cross-cohort variation within an institution (Hoxby, 2000, Figlio, 2007); natural shocks or policy-driven changes affecting peer group composition (Angrist & Lang, 2004; Gould, Lavy & Paserman, 2004; Imberman, Kugler & Sacerdote, 2012; Figlio & Ozek, 2019); admission cutoffs for schools or classrooms (Pop-Eleches & Urquiola, 2013); and experimental assignment to peer groups: see e.g. Whitmore, 2005; Duflo, Dupas & Kremer, 2011).[2]

The main findings of this literature are that *i*) academic peer effects are positive but generally small; *ii*) the size of academic peer effects depends non-linearly on students' own academic ability; and *iii*) academic peer effects vary in large and seemingly unpredictable ways across settings.

Recent empirical studies have argued that academic peer effects could be largely driven by three types of mechanisms: *i*) student effort (e.g., Kang, 2007; Brunello, De Paola & Scoppa, 2010), *ii*) group dynamics (e.g. Lavy & Schlosser, 2011; Lavy, Paserman & Schlosser, 2011; Bursztyn &

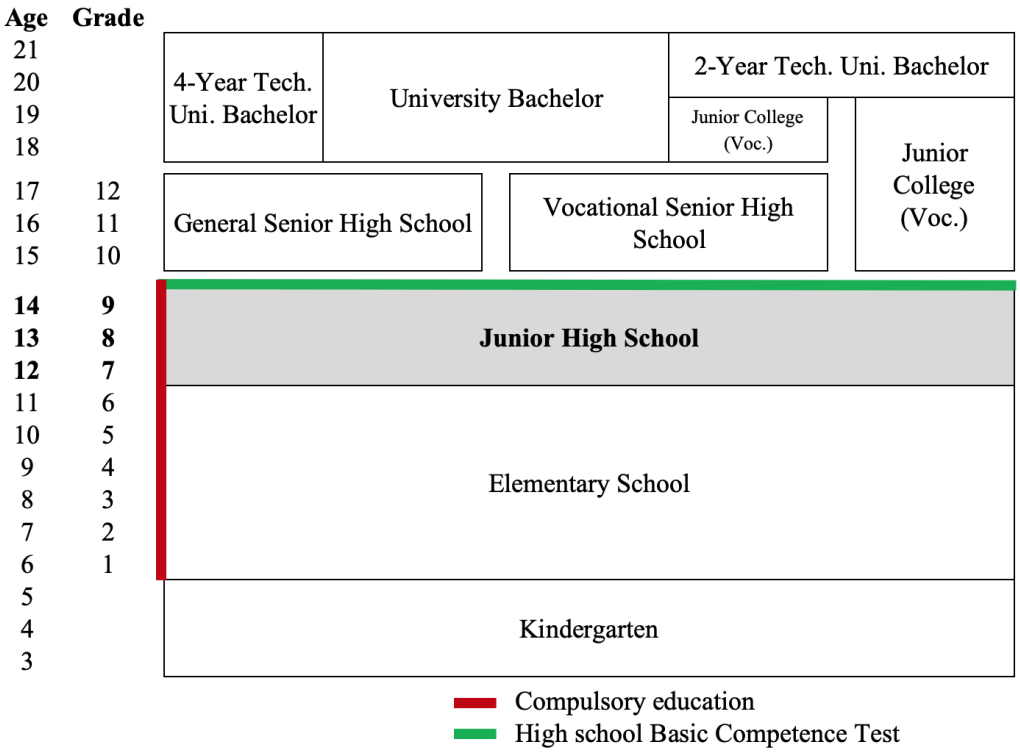
[2] It should be clear by now that there are very many studies of peer effects in education—see Sacerdote (2011, 2014) for two excellent reviews. For studies using cross-cohort variation within an institution see also: Hanushek, Kain, Markman and Rivkin 2003; McEwan, 2003; Arcidiacono and Nicholson, 2005, Hanushek, Kain, and Rivkin 2009; Lavy and Schlosser, 2011; Lavy, Paserman and Schlosser, 2011; Lavy, Silva and Weinhardt, 2012; Kiss, 2013; Diette and Uwaifo Oyelere, 2014; Kramarz, Machin and Ouazad, 2015; Gibbons and Telhaj, 2016. For studies using natural- or policy-driven shocks see also: Hoekstra, 2009; Clark, 2010; Vardardottir, 2013; Jackson, 2013; Abdulkadiroğlu, Angrist and Pathak, 2014; Dobbie and Fryer, 2014; Hoekstra, Mouganie, Wang, 2018.

Jensen, 2015; Brady, Insler & Rahmam, 2017; Feld & Zölitz, 2017), and *iii*) teacher effort or school resources (e.g., Duflo, Dupas & Kremer, 2011; Chetty et al. 2011; Hoekstra, Mouganie & Wang, 2018, Todd & Wolpin, 2018).

A separate literature, yet directly relevant for our study, emphasizes the importance of parents as drivers of their children's academic achievement. This literature models academic achievement through an education production function framework—that is, as an output produced from students', parents' and teachers' inputs and governed by well-defined production technologies (such as dynamic or technical complementarities). Recent studies in this literature show, for example, that the benefits of class size reductions are driven by changes in student effort and classroom disruption (Lazear, 2001; Finn, Pannozzo & Achilles, 2003), as well as by changes in teacher behavior (Sapelli & Illanes, 2016) and parental investments (Bonesronning, 2004, Jacob and Lefgren, 2007; Datar & Mason, 2008; Fredriksson, Oeckert, & Oosterbeek, 2016). Recent studies estimate structural models of education production functions that include school peers, parents and neighborhoods as inputs (e.g. Agostinelli, 2018, Agostinelli, Doepke, Sorrenti & Zilibotti, 2020).

In this paper, we estimate the contribution of higher-achieving academic peers to students' test scores and to many educational inputs that may also contribute to improving test scores in their own right. Conceptually, our reduced-form models map the contribution of higher-achieving peers in a linearized version of education production functions. The downside of this approach is that we do not use economic structural information to improve identification. The upside is that our models are transparent in their identifying variation, econometrically tractable, and can easily be used to quantify the share of academic peer effects explained by educational inputs via standard mediation analyses. To take full benefit of this approach we exploit the pairing of Taiwan's policy of random classroom assignment within schools and the rich data in the Taiwan Educational Panel Survey, which we describe in detail in Section 3.

Figure 1. The education system in Taiwan



3 Institutional setting & Data

3.1 Education in Taiwan

Figure 1 shows the basic organization of the Taiwanese educational system. Compulsory education in Taiwan starts at primary school, at 6 years old, and ends at the end of junior high school (middle school), around 15 years of age. In practice, however, 95 percent of students continue further onto either General or Vocational Senior High School or Junior College.

Since the democratization process in Taiwan started in the 1990s, junior high schools have been managed at the municipal level. Students can attend any school they chose but there is preferential school access based on catchment areas within each municipality. The educational curriculum is developed centrally by the Taiwanese Ministry of Education and has no subject specialization until only after junior high school. This unified curriculum is centered around sciences and mathematics

and its adoption is often cited as the reason why Taiwanese pupils are consistently placed at the top on international educational rankings (e.g. 4th out of 72 countries in PISA 2015; Law, 2005).

Critical for our identification strategy, since the 1990s municipalities are also mandated by the government to ensure the random assignment of students to classrooms within schools. This requirement was formalized by the *Implementation Guideline for Class Assignment of Junior High School Students*, later superseded by Article 12 of the *Primary and Junior High School Act* in 2004.[3] Classroom assignment plays a persistent role in students' education since students typically remain with their assigned class and homeroom teacher (or 'Dao Shi') throughout all three years of junior high school (Chang et al. 2020).

The Taiwanese education system is an extremely high pressure, exam-based, learning environment. There is a National Basic Competence Test at the end of junior high school and its results play a key role for admissions to senior high schools and senior vocational schools. A good placement in these competitive schools, in turn, results in good placements in tertiary education programs, which have high returns in the labor market afterwards. Consequently, students spend substantial time and effort preparing for these exams, and schools regularly organize practice exams and other forms of preparation. Parents are also heavily involved in their children's preparation, investing in extracurricular tutoring in mathematics, English and sciences largely through "cram schools"—private extra-curricular institutions preparing for higher education entrance examinations—throughout junior high school or even earlier.

3.2 The Taiwan Education Panel Survey (TEPS)

We use data from the Taiwanese Education Panel Survey, a project jointly funded by the Ministry of Education, the National Science Council, and the Academia Sinica. The TEPS is a nationally representative longitudinal survey of the education system in junior high school, senior high school, vocational senior high school, and junior college.

[3] Additional details can be found at: http://edu.law.moe.gov.tw/EngLawContent.aspx?id=142.

We use the junior high school sample of the TEPS. This sample includes information on more than 20,000 students, their parents, their teachers and their school administrators over two waves. The first wave was collected in early September 2001 at be beginning of students' first year of junior high school, right after their assignment to classrooms. The second wave was collected in 2003, at the beginning of the students' last year of junior high school.

Paired with the mandate of random assignment to classrooms in schools, there are three other key features of TEPS that aid our study. First, its sampling framework allows us to observe a random sample of classmates in each junior high school classroom included in the survey. TEPS follows a stratified nested sampling procedure where first 338 randomly selected junior high schools were sampled (45 percent of all high schools in the country at the time), with different sampling strata for urban and rural areas, public and private schools, and senior high and vocational schools. In each of these schools an average of three classrooms of first-year students were then randomly sampled. In each of these classrooms, around 15 students were then randomly sampled. The mandated maximum class size at the time was 35 students per class, which implies that observed students in any classroom generally represent a random half of the classroom.[4] This sampling framework is similar to that of the National Longitudinal Study of Adolescent to Adult Health (Add Health), a panel study of a nationally representative sample of middle and high school pupils in the United States. Add Health is unique in collecting friendship ties and in observing multiple cohorts of students in each school, which makes it particularly appealing for peer effect and network research (e.g. Agostinelli, 2018; Elsner & Isphording, 2017; Card & Giuliano, 2013; Bifulco, Fletcher & Ross, 2011; Calvo-Armengol, Patacchini & Zenou, 2009).

Second, and unlike Add Health, students in the TEPS take a standardized test in waves 1 and 2 called the Comprehensive Analytical Ability test. This test measures of students' cognitive ability and analytical reasoning, and it was specifically designed to capture gradual learning over time. There are 75 multiple-choice question in the test, covering general reasoning, mathematics, Chinese and English. These questions were taken from an extensive question bank which includes

[4] There are other minor sampling restrictions that are irrelevant for our empirical design, we refer the interested reader to TEPS technical reports.

adapted questions from other international standardized tests, as well as questions provided by education and field experts in Taiwan. The Comprehensive Analytical Ability test scores, constructed as the sum of all correct answers, provide excellent measures of academic ability for students and their peers.

Third, TEPS has a wealth of questions measuring student behavior, attitudes and beliefs in and outside the school environment, parent-child interactions and parental investments. TEPS also has detailed information on teachers and school administrators. We aggregate these questions to construct a truly extensive battery of measures of student, teacher and parent inputs in students' educational production function. These input measures allow us to extensively explore potential mechanisms behind academic peer effects.

To measure student, teacher and parent educational inputs in both waves, we use summative scales—sums of the answers to each question included in the scale—constructed and validated using exploratory and confirmatory factor analyses (see Appendix A for a detailed explanation and Appendix Table A.1 for summary statistics and factor loadings on all scale questions). We measure student inputs through five scales of student school effort, initiative in class, mental health, truancy, and academic self-efficacy, and three additional dummies for whether students cheat in exams, aspire to go to university, and expect to be able to go to university. We measure parental inputs though four scales of money investments, time investments, parental strictness and parental support, and three additional dummies for whether parents have conflicts with their child, use harsh punishment, and aspire for their child to go to university. Lastly, we measure school and teacher inputs though two scales of perceived quality of the school environment and of teacher engagement based on questions asked to students. Table 1 shows a high-level summary of the key academic ability and of the educational input measures we construct using the TEPS data, the number of items we can use for each measure, and the number of values each measure takes.

We use many specific pre-assignment characteristics in the TEPS data to provide evidence of random assignment of students to classrooms within schools using state-of-the-art tests in the empirical peer effects literature, which we discuss below. After excluding a few students without test scores in wave 1, our initial TEPS data includes 19,957 in 333 schools, assigned to 12,071 distinct classrooms.

3.3 Testing for random assignment in the empirical peer effects literature

A growing number of peer effects studies have relied on experimental or quasi-experimental data in which students are randomly assigned to peer groups. This literature typically uses three types of test to show that data is consistent with (conditional) random assignment of students to groups. In the first method, researchers regress student *i*'s pre-determined characteristics on the classroom leave-out mean—that is, the classroom mean after excluding student *i*—of the key regressor of

Table 1. Description of academic achievement and educational input measures

Measure	Description	Wave 1 Items \ Values	Wave 2 Items \ Values
	Students		
Test scores	Comprehensive Analytical Ability standardized test, measure of cognitive ability	75/66	75/59
School effort	Study effort; Homework on time (in English, Chinese and math class)	7 \ 23	7 \ 25
Initiative in class	Initiative to ask and answer questions (in English, Chinese and math class)	3 \ 12	3 \ 12
Cheating in exams	Student ever cheats in exams	1 \ 2	1 \ 2
Mental health	Feeling troubled, depressed, suicidal, nervous, unfocused, pressured, irritated, isolated, guilty	6 \ 19	12 \ 22
Truancy	Skipping class, fighting, watching porn, drinking alcohol, stealing, running away from home	6 \ 19	4 \ 10
Academic self-efficacy	Focus, diligence, conscientiousness, initiative, eloquence, organization, cooperation, curiosity	7 \ 22	10 \ 19
University aspirations	Student wants to go to university	1 \ 2	1 \ 2
University expectations	Student expects to be able to go to university	1 \ 2	1 \ 2
	Parents		
Money investments	Out-of-school tutoring for student: cost and intensity	2 \ 10	3 \ 10
Time investments	Frequency of going to bookstores and cultural events together with student	2 \ 7	2 \ 11
Parent-child conflict	Student quarrels with father and mother	1 \ 2	1 \ 2
Parental strictness	Father and mother's strict discipline with student	2 \ 7	2 \ 17
Parental support	Father and mother discuss future, listen carefully, worry and give advice, accept student unconditionally	8 \ 25	8 \ 7
Harsh parenting	Parents use harsh punishment with student	1 \ 2	1 \ 2
University aspirations	Parents want student to go to university	1 \ 2	1 \ 2
	Schools & Teachers		
School environment	Student perception of school study ethos, campus safety, school fairness, engagement of school administrators	5 \ 16	5 \ 16
Teacher engagement	Student perception of whether teacher knows names of students, encourages students who work hard, uses several different teaching materials, gives homework, cares about students, reviews questions after exams	6 \ 19	6 \ 36

interest. This is usually a classroom leave-out mean of ability, gender or other pre-determined student behavior (see e.g. Carrell, Sacerdote & West, 2013; Eble & Hu, 2019). A significant coefficient on the key regressor classroom leave-out mean indicates that students "treated" with peers who share a key characteristic differ from other students in the pre-determined characteristics tested. Because this test mirrors balancing-of-covariates tests in the experimental literature, we refer to them as *balancing tests*.

In the second method, researchers regress student *i*'s pre-determined characteristics on classroom leave-out mean of that same characteristic (e.g., Sacerdote, 2001). A positive coefficient on the characteristic classroom leave-out mean indicates that students are sorted into classrooms based on the characteristics tested; hence we call these *sorting tests*. Guryan, Kroft and Notowidgo (2009) observe that empirically, even under random assignment, coefficients of sorting tests present a small negative bias; they show that this small, mechanical negative correlation between own and peer characteristics seems to disappear when controlling for school-level leave-out-mean of the characteristic. Jochmans (2020) argues that Guryan, Kroft and Notowidgo's empirical correction results in low power for detecting sorting. He further derives analytical expressions for this bias in within-school estimators and proposes a bias-corrected *sorting* test which solves the power issue of previous sorting tests.

In the third method, researchers run *permutation-based sorting tests* (e.g., Carrell, Sacerdote and West, 2013; Lim and Meer, 2017). These tests go as follows. While keeping the core structure of the data (e.g., assignment to schools), researchers simulate what would happen under random assignment to treatment (e.g., to classrooms). Based on this new placebo assignment they then calculate key placebo statistics of interest—sometimes for sorting tests, sometimes for balancing tests, and sometimes for their main results. They repeat this process, say 10,000 times, and each time store their key placebo statistics. Finally, they calculate the proportion of times their placebo statistic has a more extreme value than their actual key statistic. They then calculate the proportion of times the coefficient of the placebo classroom leave-out mean is more extreme than their coefficient of the classroom leave-out mean as observed. This proportion of more extreme occurrences under placebo is a simulation-driven empirical p-value for a test of random assignment and can be judged by typical standards of statistical significance. These empirical p-values could be calculated for many statistics of interest, including for sorting and balancing tests but also for

such tests at the school or even classroom level. When many of these empirical p-values are calculated, researchers can aggregate them into one overarching statistical test using goodness-of-fit tests for the distribution of p-values, which should be standard uniform under random assignment to treatment.

All three methods above are valid ways to produce evidence of quasi-random assignment, yet all methods also have their shortcomings. Neither method naturally corrects for multiple testing when researchers use many pre-determined characteristics in their tests. Using multiple hypotheses testing corrections (e.g., Benjamini & Hochberg, 1995; Romano & Wolf, 2005) can, in turn, severely decrease test power. Another approach is to joint-test the significance of all pre-determined characteristics in predicting treatment but these joint tests have a tendency to over-reject, especially when using cluster-robust inference methods (Pei, Pischke & Schwandt, 2018). Permutation tests have the additional problem of being relatively complex to program since researchers are required to keep most of the data structure identical (e.g., assignment to schools, number of classrooms in each school, class size) while still reassigning treatment at random, then correctly recalculate all treatment measures, and ensure that treatment variation is correctly accounted for in all estimates – which is harder with discrete measures of pre-assignment characteristics like gender or race. In addition, goodness-of-fit tests used to aggregate many empirical p-values in permutation tests, such as the Kolmogorov-Smirnoff, have known power issues.

Given the volume of peer effect studies out there, it is no surprise that in many of them there is evidence of some systematic assignment to peer groups (e.g. Krueger, 1999; Krueger & Whitmore, 2001; Whitmore, 2005; Dee, 2004; Ammermueller & Pischke, 2009; Balsa, Gandelman & Roldán, 2018). When tests of random assignment reject the null that students are randomly assignment to peer groups, researchers have used three types of econometric strategies.

A first approach is to adapt the econometric specification and adjust the interpretation of estimates accordingly (Krueger 1999, for example, estimates intent-to-treat effects rather than treatment effects), or to consider the size of the selection bias when interpreting results (e.g. Dee, 2004). This can be appropriate if the evidence on systematic assignment is weak, quantitatively small, and does

not hint at further systematic assignment based on unobservable characteristics that affect student outcomes. The cost, however, is that estimates might be biased if any of these conditions fail.

A second approach is to remove treatment clusters where the data are consistent with some form of systematic assignment to treatment (e.g. Krueger, 1999; Whitmore, 2005; Chetty et al., 2011). This approach is valid if there are clear reasons to believe that random assignment applies to some *known* treatment clusters but not others, which usually requires intimate knowledge of the institutional background behind the data and the presence of markers of these known clusters. In complex institutional settings, removing data clusters suspected of systematic assignment to treatment quickly becomes unfeasible: depending on the number of potential clusters with systematic assignment, how similar they are to one another, and whether the data contains clear markers for these clusters, excluding them from main analyses can be costly in terms of statistical power.

A third approach is to control for pre-assignment characteristics that reveal systematic assignment in the preferred specification, thus relying on mean independence of treatment *conditional on these characteristics* (e.g., Lavy, Paserman & Schlosser, 2011; Gong, Lu & Song, 2019). This approach is not costly in terms of power and does not require intimate knowledge of the institutional background, yet it assumes (often implicitly) that controlling for characteristics related to systematic assignment fully accounts for related unobserved characteristics that also determine assignment. Economist are often wary of this assumption. This third approach also comes with other shortcomings. In particular, it assumes that a single parameter function (e.g., linear) in the pre-assignment characteristics is sufficient to account for systematic assignment. This assumption is unlikely to hold if there are several such characteristics or several treatment clusters that differ in their drivers of systematic assignment. Parametrically relaxing this assumption can quickly become costly in terms of power. Perhaps more importantly, controlling for pre-assignment characteristics changes the interpretation of the peer effect estimates, often making them less immediately available for designing better peer group assignment policies. For example, unbiased peer effect estimates that control for parental education can only be used to predict outcomes of reassignment policies that hold parental education constant—a difficult exercise unless student reassignment to classrooms is done explicitly on parental education, which is unlikely to happen in practice.

In sum, there are several ways to test for random assignment of students to peer groups and several ways to deal with an eventual rejection of random assignment. None of the tests are perfect, nor are the solutions. In the next section, we show our main test for random assignment in the TEPS and refer the interested reader to Sections 5.1.1 and 5.1.2 for the additional tests we run. The case of the TEPS also presents an interesting challenge that combines *i*) a national mandate of random assignment of students to classrooms within schools, *ii*) incentives for parents and schools to violate this mandate *if they believe that higher-achieving peers affect student outcomes*, and *iii*) unusually rich pre-assignment data to test the outcome of these two clashing institutional features. In the next section, we also propose a new data-driven method for finding subsamples where quasi-random assignment is credible, which is particularly useful in complex institutional settings such as ours.

3.4 Random class assignment in TEPS and balancing on peer ability

Our identification strategy exploits random assignment of students to classrooms. If random assignment holds, we expect our treatment of interest, classroom leave-out-mean of peer ability, to be randomly assigned to students. Random assignment *to treatment* is the main identification assumption under which our coefficient estimate yields a causal estimate of the effect of peer ability on subsequent outcomes.

To show that the data are consistent with random assignment to classrooms within schools, we run sorting tests in the complete TEPS data on standardized test scores and 17 pre-assignment characteristics. We start from the complete sample, to prevent missing values to lead to over-rejecting sorting tests of random assignment. In this complete data, we find evidence of sorting by student ability and by several other student characteristics. We take this as evidence that students are not randomly assigned to classrooms across the entire TEPS data (see Appendix Table C.1).

There are many reasons why, in defiance of the national mandate of random assignment, we could find evidence of systematic assignment of students to classrooms. These can range from school principals occasionally catering to some parents' preferences for their child to be assigned to some classrooms, to institutionally allowed "talent" classrooms that pool high-ability student together, to a more concerning blatant disregard for the national mandate across schools. We develop a data-driven procedure that helps us determine the reason behind this seeming violation of random

assignment in the data, and identify a sample where random assignment likely holds. We describe the key features of this procedure below, and refer the interested reader to a more complete description in Appendix C.

The Fishing Algorithm

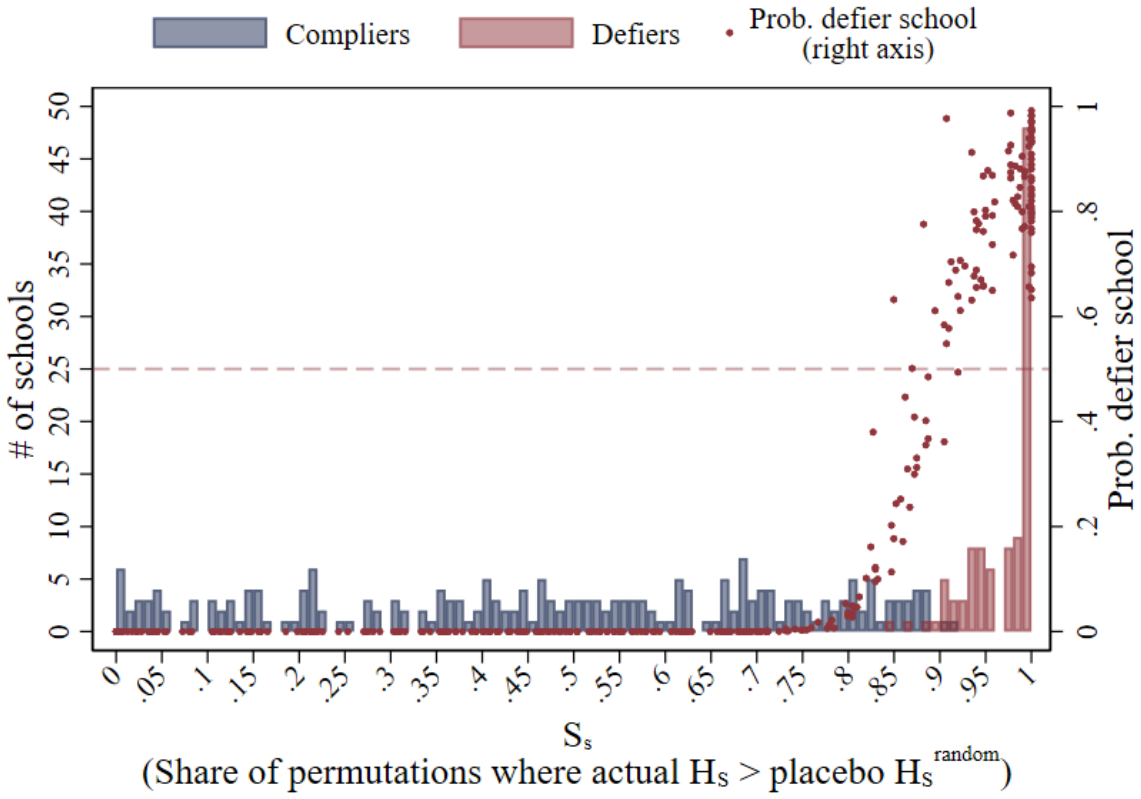
Since the law in Taiwan has an explicit mandate of random assignment of students to classrooms, we suspect that rejecting the null of sorting tests is most likely driven by few "defier" schools that systematically sort students. Unfortunately, our data does not allow us to infer directly which are these schools to exclude them from our analysis.

We therefore designed a sample trimming method, which draws from randomization inference insights and combines clearly pre-defined selection rules and latent-class modeling. Our "Fishing Algorithm" is a data-driven approach to identify and exclude the schools that show evidence inconsistent with conditional random assignment. Given the goal of this paper, we focus on trimming schools that systematically sort students of similar academic ability into classrooms, although our method can be easily adapted to trim schools that sort on any observed characteristic in the data, and even on multiple characteristics at once.

The key five steps of our fishing algorithm are the following. First, we construct for each school a measure of *strength of sorting*; how strongly is the school sorting students of similar ability into the same classrooms. This measure is akin to a Herfindahl-Hirschman index of ability concentration in classrooms within each school, with larger values indicating stronger ability sorting in classrooms in the school. We call this measure H_s. Second, for each school we use several permutations of randomly assignment students to classrooms within school without replacement and construct, for each simulated classroom assignment, its corresponding simulated index H_s^{random}. This procedure recovers the distribution of ability concentration in classrooms for each school under the null of random assignment. Third, for each school we compute the share of permutations for which the observed index H_s in the data was larger than the simulated index H_s^{random} under the null of random assignment, and call this share S_s. Under perfect compliance

with random assignment, we expect the school distribution of S_s to be uniform over the [0,1] interval; if random assignment is more values of S_s close to 1 and therefore a less uniform distribution. Fourth, we estimate the latent probability that each school is a "defier" schools (i.e., a school that sorts students into classrooms more strongly than chance would allow) using latent class modelling—an atheoretical data-driven partitioning method that finds observations (e.g., school shares S_s) that are likely to be generated by a the same stochastic process (e.g., ability-sorted classroom assignment). Using school-level data, we fit a finite mixture model where the outcome is S_s, the regressors are constants for each latent class, and we include school-level variables that could help identify defier schools (such as the share of parents who report pushing to get their children assigned to a better classroom). One or more of the latent classes in this model

Figure 2. Schools identified as defying random assignment using our fishing algorithm



This figure shows the school-level distribution of our measure for whether schools sort students into classrooms more strongly than chance would allow, given the school size, number and classroom size and student composition. The probability of being a defier school is the posterior probability of being in a latent class classified as defiers by us and calculated based on a finite mixture model of school sorting using several school averages of parental characteristics as class predictors. See Appendix C for details.

correspond to schools with improbably high S_s—the likely defier schools—and the model itself produces school-level posterior probabilities of each school belonging to this defier class. In the fifth and final step, we flag defier schools based on whether their posterior probability of being in the defier class is larger than 0.5. As mentioned above, a more complete description of this fishing algorithm can be found in Appendix C, and we provide validation of this procedure using simulated data in Appendix D.

Most schools in the TEPS data show evidence consistent with random assignment, whereas some schools present obvious evidence of sorting (Appendix Figure C1). As illustrated in Figure 2, our fishing algorithm identifies 106 out of the 333 schools in TEPS as defier schools, which we exclude out of our estimation sample. This leaves us with a trimmed sample of 13,685 students in this

schools, allocated to 853 classrooms (68 percent of the TEPS data). Our trimmed sample is very similar to the overall TEPS data in terms of all key student and parent characteristics in wave 1, and is also similar to our final estimation sample of 11,068 observations with complete information on student and peer test scores and educational inputs (Appendix Table C2).

An important concern in applying our fishing algorithm is over-trimming; that is, to remove schools that by chance look like defers but are not. This process will almost always result in some schools being over-trimmed. Over-trimmed schools would have contributed useful variation to identify peer effects. With severe over-trimming, peer effects could be less precisely estimated at best, and biased at worst (upwards if e.g., peer effects are highly non-linearly driven by the positive effect of higher-achieving peers on high-achieving students). However, over-trimming is also easily diagnosed in our algorithm; it shows as *negative and significant* post-trimming sorting t-statistics. If negative post-trimming sorting t-statistics occurs, researchers should make efforts to improve the performance of the fishing algorithm (by e.g., finding better predictors of sorter schools or exploring different latent lass structures or models). If no improvement can be made, it is important to highlight the over-trimming brought on by the algorithm and cautiously interpret findings accordingly. Fortunately, in our application of the fishing algorithm to TEPS we find virtually no evidence of over-trimming.

Sorting and balancing tests in our trimmed sample

Table 2 presents the results of sorting and balancing tests on the trimmed sample, that is, once we exclude the schools likely to be non-compliant with the mandate of random assignment.[5] Columns 2 and 3 show sorting tests t-statistics, to be compared to standard normal critical values, whereas

[5] For this discussion, it is useful to keep in mind the omitted variable bias formula for our peer effect estimator β:

$$E[\beta|X] - \beta = \gamma\rho$$

where γ is the conditional effect of any omitted factor on student outcomes and ρ is proportional to the correlation between the omitted factor and our peer achievement leave-out-mean. Evaluating all endogeneity concerns against this formula is an enlightening way to map econometric endogeneity concerns to economic principles.

columns 3 and 4 show coefficients and standard errors of balancing regressions of pre-assignment characteristics on peer ability.

Table 2. Balancing and Sorting Tests on our Trimmed Sample

Treatment:		Sorting tests (t-statistic)		Balancing tests	
		Peer outcome leave-out-mean		Peer ability leave-out-mean [std]	
	Students	Guryan et al. (2009)	Jochmans (2020)	Coef.	Std. err.
Outcomes: Pre-assignment characteristics					
Student test scores [std]	13,685	-0.2	0.1		
Female student	13,685	2.1	-0.2	0.008	(0.011)
Student born before 1989	13,611	-0.8	0.6	-0.005	(0.010)
Household income > NT$100k/mo.	13,454	-0.7	-0.3	-0.019***	(0.007)
College-educated parent(s)	13,084	-0.8	0.8	0.001	(0.009)
Parent(s) work in government	13,023	1.4	0.0	0.010	(0.007)
Ethnic minority parent(s)	13,081	2.2	1.4	-0.004	(0.009)
Prioritized studies since primary school	13,593	-1.7	0.8	-0.010	(0.009)
Reviews lessons since primary school	13,583	-0.2	1.7	0.003	(0.008)
Likes new things since primary school	13,554	1.5	2.4	-0.001	(0.011)
Was truant in primary school	13,489	1.6	-0.7	0.000	(0.011)
Student had mental health issues in primary school	13,486	-0.7	0.2	-0.004	(0.010)
Had private tutoring before junior high	13,525	0.3	1.4	0.004	(0.012)
Family help with homework before junior high	13,013	1.2	0.8	-0.020**	(0.008)
Student quarreled with parents in primary school	13,502	-1.5	-1.2	-0.001	(0.009)
Student enrolled in gifted academic class	13,554	-1.2	1.8	0.013	(0.008)
Student enrolled in arts gifted class	13,554	2.2	2.9	-0.013	(0.015)
Parents made efforts to place student in better class	13,508	2.2	3.2	0.035***	(0.010)

*Estimates in our trimmed sample of 227 schools and 853 classrooms. All estimators include school fixed effects. The reference distribution for the Guryan et al. (2009) and the Jochmans (2020) sorting statistics is the standard normal. The last column reports cluster-robust standard errors at the classroom level. ***, ** and * mark estimates statistically different from zero at the 90, 95 and 99 percent confidence level.*

The main endogeneity concern in our estimates is ability sorting of students; that is, that high-ability students are assigned together in the same classroom. This type of sorting is concerning because, if ability is dynamically self-productive as in e.g., Cunha and Heckman (2007), it would bias peer effect estimates upwards. The first row of Table 2 shows that this sorting is not a concern in our trimmed sample.

Another common endogeneity concern is whether students are sorted in productive characteristics other than ability, say parental income. This kind of sorting is tested in the second and third columns, second row and below, of Table 2. Sorting on parental income can introduce bias in peer

effects estimates if these characteristics are related to student achievement. Note, however, that if income sorting were related to students' achievement at baseline, this sorting would have already been reflected in the baseline achievement sorting. This still leaves the possibility that parental income has not been productive for student achievement *at baseline* but might become productive afterwards. If that is the case, income sorting at baseline can bias peer effects upwards over and above achievement sorting. There is not much evidence of sorting on other characteristics in our trimmed sample, especially when using the Jochmans (2020) state-of-the-art test. There is some evidence of sorting on intellectual curiosity and, perhaps more importantly, sorting for students enrolled in gifted arts classrooms and students whose parents report making efforts to get them assigned to particular classrooms. Several institutional settings, including TEPS, could allow for this type of sorting to occur over and above achievement sorting.

For student sorting on other characteristics to introduce bias in our peer effect estimates, however, a second necessary condition is for the student characteristic to be related to our peer achievement leave-out-mean measure. The last two columns of Table 2 show these tests. In our trimmed sample, the only potentially concerning characteristic which *i*) could affect student achievement over and above baseline achievement, *ii*) students are sorted on at baseline, and *iii*) is also related to peer achievement at baseline is whether parents made efforts to get their child assigned to a particular classroom. Of all the other characteristics that we test, only family income and family engagement with homework before baseline are *negatively* related to peer achievement. This last finding likely happens because trimming schools that sort based on achievement will reduce the relation between peer achievement and characteristics correlated to student achievement (such as income and homework engagement). These negative correlations with peer achievement should not bias peer effect estimates once we control for students' own ability, which we do in all our models. Still, to account for any leftover correlated selection we also include controls for household income, family engagement with homework, gifted art classroom assignment, and parents' pushiness to get child assigned to a particular classroom in our main specifications. We jointly refer to these as balancing

controls and note that they are neither crucial for our empirical design nor do they affect any of our results.[6]

Overall, our fishing algorithm is an effective way to identify schools that systematically assign student to classrooms in our data. In the schools identified by the algorithm as balanced we find no substantive evidence of systematic assignment, and we will keep this trimmed sample as our estimation sample throughout our main analyses. In Section 5 we also show the results of a battery of additional sorting tests, discuss in detail other ways to identify our estimates, explore the issues of sample selectivity, and compare our trimmed sample with the initial TEPS sample.

4 Main results

4.1 The effect of classroom peer test scores on own test scores

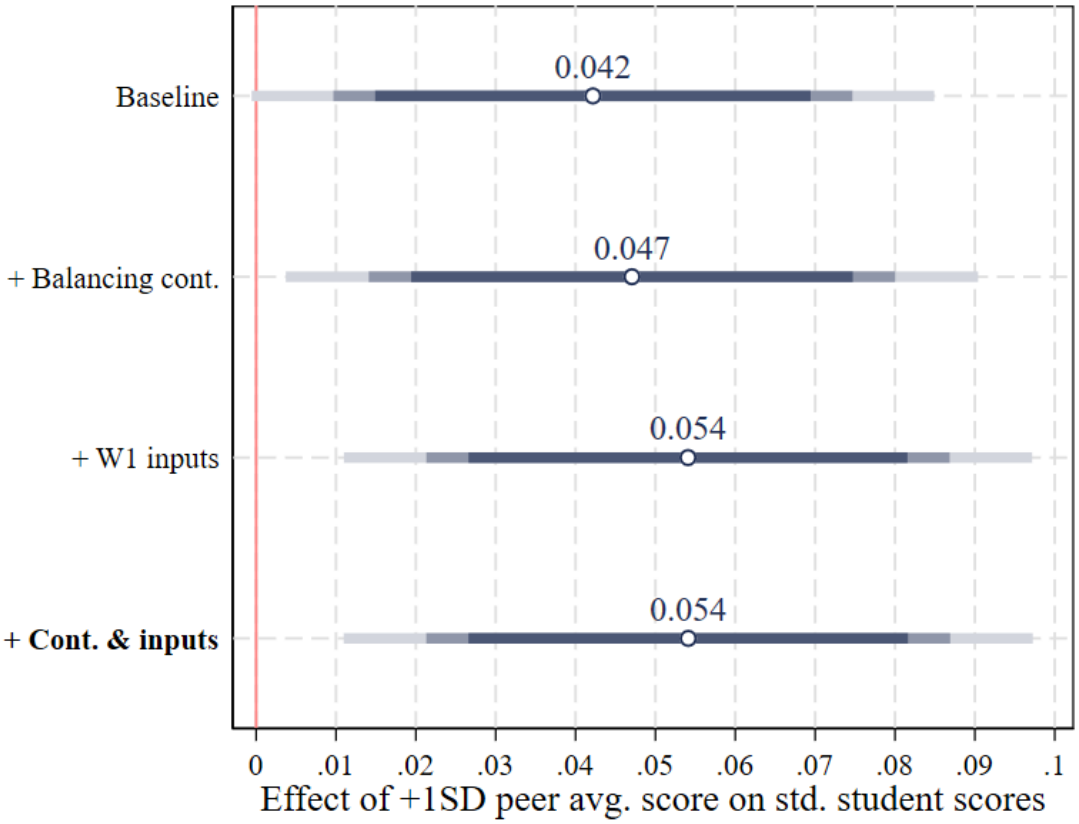
Now that we have established a sample where conditional random assignment of students to classrooms holds, we go on to establish the existence of academic peer effects. In its most basic form, we do this by regressing students' standardized test scores in wave 2, $\text{Test Score}_{\text{ics2}}$, on the standardized classroom leave-out mean of test scores in wave 1, $\overline{\text{Test Scores}}_{\text{ics1}}^{-\text{i}}$, our measure of average peer test scores. To this simplest specification we add school fixed effects and students' own test scores in wave 1, both crucial for identification (Angrist, 2014). We also consider specifications with and without the additional balancing controls (household income, family engagement with homework, gifted art classroom assignment, and parents' pushiness to get child assigned to a particular classroom) and standardized scales of student inputs (school effort, initiative in class, truancy, academic self-efficacy, and mental health), parent inputs (investment in private tutoring, time investments, parental strictness and parental support), school and teacher inputs (school environment and teacher engagement). We do this to assess the extent to which

[6] Also, note that due to the power in our data, we detect small differences in balancing tests that would have likely gone unnoticed in other designs. Our ex-post Minimum Detectable Effects (MDEs) for our balancing tests are as small as 2.2 percentage points in the chance of being female, and less than 1 percentage point in the likelihood of having a migrant background. For comparison, the MDEs of balancing tests are 17 percent of a standard deviation in math test scores in the STAR data (Dee, 2004), and 25 percentage points for being female and 10 percentage points for migrant in the Add Health data (Bifulco, Oh and Fletcher, 2014).

these covariates could capture omitted variable bias in our peer effect estimates. We cluster standard errors at the classroom level.

Figure 3 shows strong positive peer effects in our setting. It further shows that including balancing controls or wave 1 inputs does not qualitatively change our estimates, though it does slightly increase precision. This estimate stability is a reassuring result which provides strong evidence of no omitted variable bias in our estimates, especially given the wide range of controls included in our educational input measures.

Figure 3. The effect of better peer test scores on students' own test scores in wave 2



This figure reports estimates of regressing standardized student test scores in wave 2 on standardized average peer test scores in wave 1 in our sample containing 227 schools, 853 classes, and up to 12,816 students. Rows present results of models with different sets of control variables. The Baseline model includes wave 1 student test scores and school fixed effects. Balancing controls include household income, family engagement with homework, gifted art class assignment, and parents' efforts to get child assigned to a particular classroom. W1 inputs include standardized scales of student inputs (school effort, initiative in class, truancy, academic self-efficacy, and mental health), parent inputs (investment in private tutoring, time investments, parental strictness and parental support), school and teacher inputs (school environment and teacher engagement). Horizontal bars `show the 99%, 95% and 90% confidence intervals for each estimate, based on standard errors clustered at the classroom level. Estimates in this figure are also shown in Appendix Table B1.

Our preferred specification is on the last row of Figure 3, highlighted in bold. This specification controls for school fixed effects and student wave 1 test scores, as well as all wave 1 educational inputs and our four balancing covariates. It therefore identifies academic peer effects within Todd and Wolpin's (2003) cumulative value-added specification; holding constant past outputs and educational inputs. This will prove important in the following sections. Our preferred estimates can be re-expressed as:

$$\text{Test Score}_{\text{ics2}} = \underset{(0.017)}{0.054}\,\overline{\text{Test Scores}}^{-\text{i}}_{\text{ics1}} + \underset{(0.009)}{0.562}\,\text{Test Score}_{\text{ics1}} + \hat{\theta}'\text{Controls}_{\text{ics1}} + \hat{\mu}_{\text{s}} \qquad (1)$$

where $\text{Controls}_{\text{ics1}}$ includes balancing controls and wave 1 educational inputs.

These estimates imply that having one standard deviation higher average peer test scores in wave 1 increase own test scores by 5.4 percent of a standard deviation in wave 2. Comparing effect sizes in this literature is quite difficult; differences in standardized effect sizes across studies could capture true differences in responses to peer ability but could also reflect differences in standard deviations in peer achievement and student outcomes across settings. Assuming these standard deviations are comparable across studies, our peer effects are also similar (e.g. Imberman, Kugler and Sacerdote 2012; Brunello, De Paola and Scoppa 2010). Compared to studies where students are randomly assigned to peer groups, our estimates are around the median of estimate. Yet our estimated effect measures the impact of two years' worth of exposure to classroom peers, which represents a strong dose compare to most comparable studies, thus our effect could also be seen as relatively small.

To give this number more perspective, our estimated effect of a 1SD increase in average peer scores is about a tenth of the estimated effect of a 1SD increase in students' own lagged test scores. Our peer effect estimate is about half the marginal effect of having at least one college-educated parent, and about a sixth of the unconditional test score gap between children of two-parent households and single-parent households.

Another way of sizing the impact of higher-achieving peers is through the lens of socioeconomic inequality. Due largely to school sorting, the peers of poor students (with household monthly incomes under NT$20,000, corresponding to the 10th percentile poorest in the sample) have 68 percent of a standard deviation lower scores than the peers of rich students (with household

monthly incomes over NT$100,000, corresponding to the top 15th percentile). The rich-poor test score gap in wave 2 test scores gap is 1.1 standard deviations. Putting these two numbers together, our linear peer effects imply that 3.5 percent of the rich-poor gap in standardized test scores can be explained by the richer students' access to higher-achieving peers.

4.2 The effect of classroom peer test scores on educational inputs

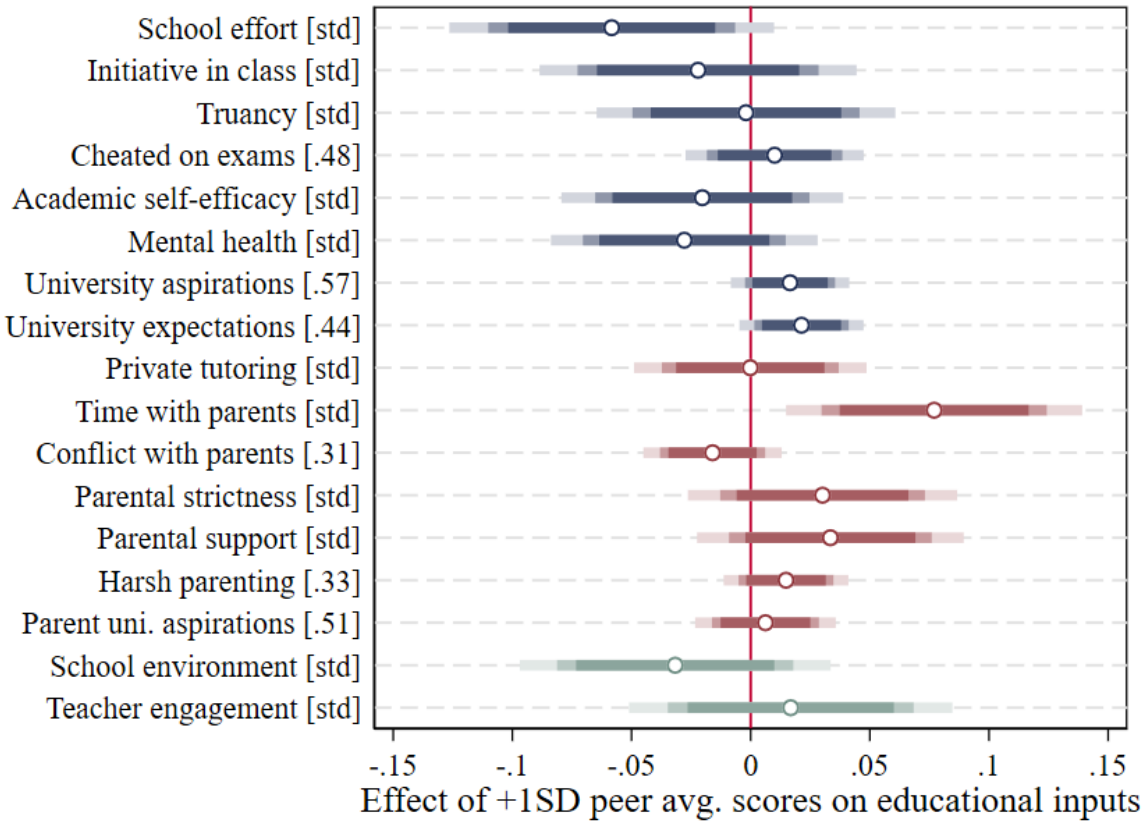
In this section, we estimate the impact of higher-achieving academic peers on educational inputs in order to explain how academic peer effects work. We estimate variations of Equation (2) using our measures of educational inputs in wave 2 as outcomes. Figure 4 shows the effect of a 1SD increase in average peer test scores on wave 2 educational inputs in our estimation sample. Each row shows the effect of peer test scores on a different educational input. We show the unconditional mean of each outcome in square brackets to give context to these estimates. Navy blue estimates show effects student inputs, maroon estimates show effects on parent inputs, and teal estimates show effects on school and teacher inputs.

A 1SD increase in average peer test scores decreases students' school effort in wave 2 by 5.8 percent of a standard deviation. While these effects are a priori surprising, they are difficult to benchmark against previous findings. Many studies have hypothesized study effort to be a key mechanism through which peer effects operate, yet few of them provide estimates of effort responses to higher-achieving peers. Among the few studies that do, there seems to be no consensus (Feld & Zölitz, 2017; Mehta, Stinebrickner & Stinebrickner, 2019; Fang & Wan, 2020).

A 1SD increase in average peer test scores also increases students' aspirations to go to university by 1.7 percentage points, and their expectations of actually going to university by 2.1 percentage points. These seem like small effects, corresponding to around 3-5 percent of their respective unconditional means, but become more sizeable when compared to the effect of other known shifters of aspirations. One could compare them, for example, to the 8.5 percent increase in parents' higher education aspirations for girls from opening access to male-dominated professions in India (Beaman et al. 2012), the 5.2 percent increase in educational aspirations of cast-priming in high-casts in India as well (Mukherjee, 2015), or the precisely-estimated null effect of university information on educational aspirations of Colombian students (Bonilla-Mejia et al. 2019).

A 1SD higher-achieving classroom peers also increase parents' time investment by 7.7 percent of a standard deviation. Our time investment measure in TEPS focuses in dinner time spent with parents, yet our estimated peer effect could be compared to half of Fredriksson, Oeckert, and Oosterbeek's (2016) impact of having one student more in one's classroom on parents' likelihood of helping the child with homework, or with a fifth of Pop-Eleches and Urquiola's (2013) effect of a child attending a marginally worse school.

Figure 4. The effect of better peer test scores on educational inputs in wave 2



This figure reports estimates of regressing educational input measures in wave 2 on standardized average peer test scores in wave 1 in our sample containing 227 schools, 853 classes, and 12,816 students. Rows present results of models with different educational inputs as outcomes. Unconditional means of each outcome are shown in square brackets, and [std] marks outcomes that have been standardized to have a mean of zero and a standard deviation of one. All models control for school fixed effects, student test scores in wave 1, balancing controls, and educational inputs in wave 1. Student, parent, and school & teacher inputs are shown in navy blue, maroon, and teal. Horizontal bars show the 99%, 95% and 90% confidence intervals for each estimate, based on standard errors clustered at the classroom level. Estimates in this figure are also shown in Appendix Table B2.

Finally, Figure 4 shows that we cannot detect effects of higher-achieving peers on many educational inputs that have previously been considered as key potential mechanisms behind peer effects, such as student initiative in class and class disruption. We estimate precise null effects on

all measures of parental investment or parenting behavior other than parental time investments. This finding is important because while we find no parental behavioral responses to classroom peer ability, previous studies have shown evidence of parental behavioral responses to other types of public investments such as school admissions or classroom size. Lastly, we also find precisely estimated null effects for additional potential mechanisms, in contrast with studies which have found suggestive evidence on students' perception about their school environment (e.g. Feld and Zölitz 2017) or teacher engagement in the classroom (Lazear, 2001; Duflo, Dupas and Kremer, 2011; Golsteyn, Non and Zölitz, *forthcoming*).

Importantly, we can detect relatively small effects for most of these mechanisms. Between all our estimates, the largest standard error for a standardized educational input is 0.026. A standard ex-post Minimum Detectable Effect (MDEs) size calculation with 95 percent confidence and 80 percent power implies that we could have detected effects as small as $0.026 \times 2.8 = 7.3$ percent of a standard deviation for outcomes such as initiative in class or teacher engagement. A 7.3 percent of a standard deviation in an outcome is a relatively small detectable effect; close to 10 percent of the gender gap in effort (women pay more effort than men), 18 percent of the difference between private tutoring investments of top-income parents and the rest, or 9 percent of the difference between the time investments of two-person and single-parent households.

Overall, we show that higher-achieving peers decrease student effort, increase student aspirations and expectations to attend university, an increase in parental time investments. We can be made sense of the first two, seemingly contradicting, results in the lens of existing theories of performance under uncertainty; they could be consistent with exposure to higher achieving peers as a form of relative performance feedback. The sign of these estimates is in line with the theoretical model and recent field evidence of Azmat et al. (2019). The latter result on time investments provide new insights on the relatively thin evidence base on parents' behavioral responses to school inputs. Our effects suggest that parents *complement* school inputs (i.e., better school peers) by increasing their own time investment. This collides with evidence that parents tend to treat school inputs and own time investments as substitutes (Pop-Eleches & Urquiola, 2013; Fredriksson, Oeckert, & Oosterbeek, 2016) but is consistent with other evidence from Taiwan that showing that parents complement teacher qualifications with financial investments of their own (Chan, Cobb-Clark & Salamanca, 2020).

More relevant is that—depending on the productivity of these educational inputs for student achievement—these input responses could all be legitimate mechanisms for explaining our 5.4 percent of a standard deviation effect of higher-achieving peers on test scores. In the next section, we calculate how much of our estimated academic peer effect can be explained by these mechanisms.

4.3 The share of the academic peer effect explained by changes in educational inputs

We are now able to formally ask how much of the 5.4 percent effect of higher-achieving peers on students' test scores can be explained by their intermediate impact on educational inputs. To do this we follow Gelbach's (2016) decomposition, which we adapt to our setting in order to use only within-school variation by modifying the `b1x2` Stata package.

This decomposition calculates the total mediated effect (ME) of educational inputs on peer effects:

$$\mathrm{ME} = \sum_{\mathrm{k}} \mathrm{ME}_{\mathrm{k}} = \sum_{\mathrm{k}} \underbrace{\frac{\partial \mathrm{Ed.\,Input}^{\mathrm{k}}_{\mathrm{ics2}}}{\partial \overline{\mathrm{Test\ Scores}}^{-\mathrm{i}}_{\mathrm{ics1}}}}_{\text{(A)}} \times \underbrace{\frac{\partial \mathrm{Test\ Score}_{\mathrm{ics2}}}{\partial \mathrm{Ed.\,Input}^{\mathrm{k}}_{\mathrm{ics2}}}}_{\text{(B)}}, \tag{2}$$

where $\mathrm{Ed.\,Input}^{\mathrm{k}}_{\mathrm{ics2}}$ stands for educational input k in our set of inputs. The terms (A) are the causal effects of higher-achieving peers in wave 1 on educational inputs in wave 2 as shown in Figure 4. The only remaining pieces for the calculation of ME are therefore the terms (B) which are the partial returns (i.e., holding other inputs constant) to each of the educational inputs on student scores in wave 2.

There is no ideal experiment for estimating (B), not even by independently and randomly varying each educational input over a period of two years and then estimating their causal impact on student test scores. The reason, as expressed by Todd and Wolpin (2003), is that such experiments would identify "policy parameters"—effect identified out of variation not subject to choices of parents or schools but exogenously induced—rather than "production function" parameters. Policy parameters are identified by variation in inputs exogenously pressed onto people, rather than by naturally-occurring variation through people's investment decisions across the population (see e.g., Imai, Tingley & Yamamoto, 2013; Keele, Tingley & Yamamoto, 2015). Thus, policy

parameters answer many important questions but they do not recover returns to inputs, so their use is limited in a mediation analysis as described by equation (2).

Todd and Wolpin (2003) argue for using (cumulative) value-added models to estimate the (B) term of equation (2). Todd and Wolpin (2007) and Fiorini and Keane (2014), among others, discuss these models in detail and show that they can identify the returns to educational inputs under relatively weak conditions, and we find ourselves in an ideal scenario for estimating these models. This is because in our setting we *i*) always use within-school variation which accounts for unobserved school-level heterogeneity, *ii*) can control for standardized test scores in wave 1, *iii*) can control for a myriad of educational inputs in wave 1, and *iv*) only need to estimate returns over a two-year period. For all these reasons, we estimate the terms (B) as the $\hat{\beta}_k$ from the within-school cumulative value-added model:

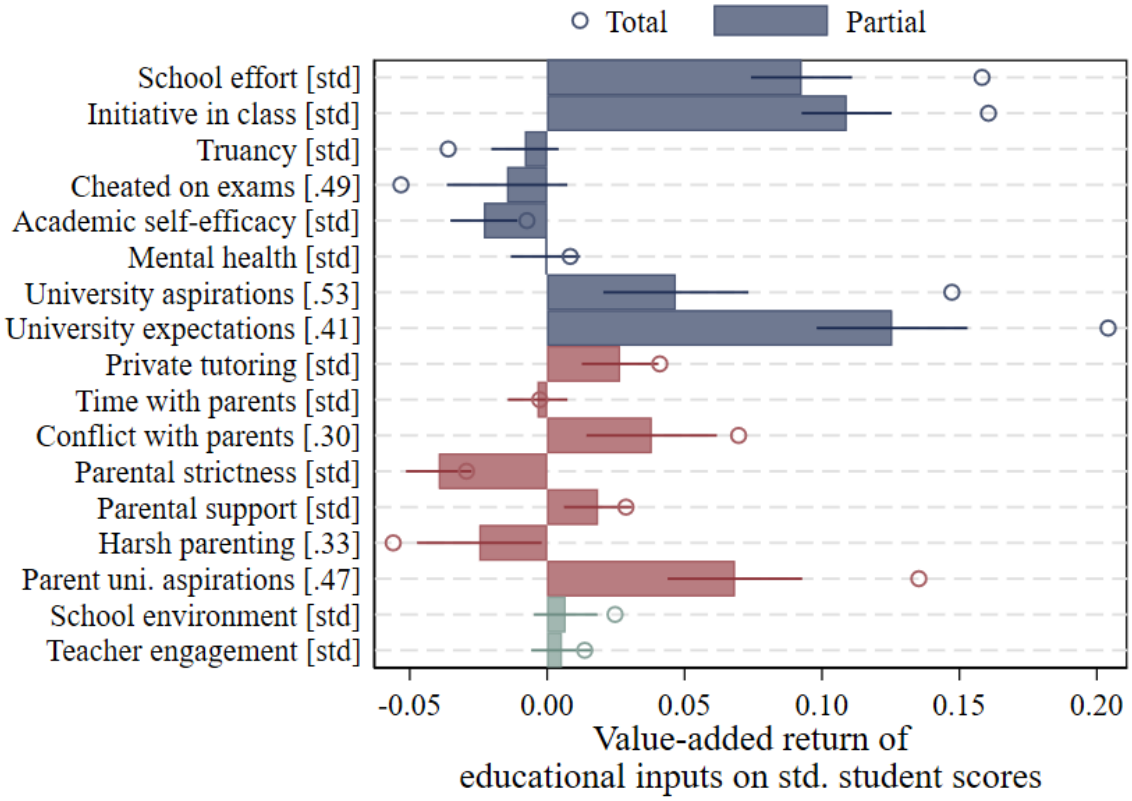
$$\text{Test Score}_{ics2} = \sum_{k=1}^{K} \beta_k \, \text{Ed. Input}^k_{ics2} + \delta \text{Covariates}_{ics1} + \gamma_s + \upsilon_{ics2}, \tag{3}$$

where Covariates_{ics1} includes student test scores, average peer test scores, and all other educational inputs in wave 1. To the extent that our school fixed effects account for school-level unobserved heterogeneity γ_s and extensive set of high-quality covariates account for endogeneity in observable educational inputs, equation (3) will identify unbiased estimates of the average partial return to each of the K educational input in our data.

Figure 5 shows the within-school cumulative value-added estimates of the total and partial average returns of educational input in wave 2. Total effects are return parameters estimated one input at the time. Partial effects are the return parameters estimates $\hat{\beta}_k$ obtained from equation (3) with the complete set of K inputs include as regressors together. In other words, they are the returns of each educational input k holdings constant all other $K-1$ inputs. We rescale test scores and all continuous inputs in wave 2 so that each value can easily be interpreted as the return of a one standard deviation increase in standard deviations of scores. The circles show the total returns of each input, and the bars show the partial effect of each input with their corresponding 95 percent confidence interval.

We obtain precise estimates of the average partial returns to all educational inputs. The first row in Figure 5, for example, shows that a 1SD increase in school effort between waves 1 and 2 carries

Figure 5. Returns to educational inputs from cumulative value-added models

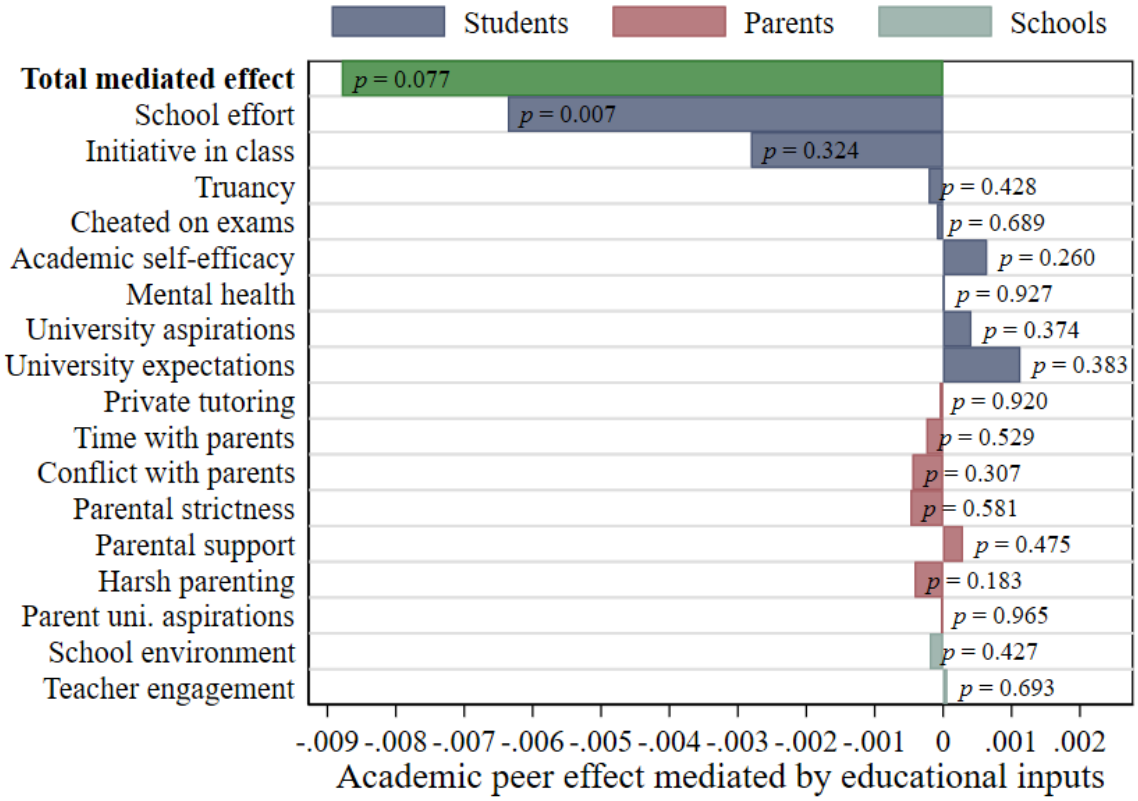


This figure reports coefficient estimates of regressing student test scores in wave 2 on educational inputs in wave 2 in our estimation sample containing 227 schools, 853 classes, and 12,816 students. Rows present coefficients of different regressors; Unconditional means of each input are shown in square brackets and [std] marks inputs that have been standardized to have a mean of zero and a standard deviation of one; circles show total effects (one input at the time) and bars represent partial effects (all inputs jointly). All models control for school fixed effects, student test scores, average peer test scores, and educational inputs in wave 1. Student, parent, and school & teacher inputs are shown in navy blue, maroon, and teal. Spikes show 95% confidence intervals on partial effects based on standard errors clustered at the classroom level. These results are also available in Appendix Table B3.

an average return of 9.3 percent of a standard deviation in test scores in wave 2. There are also positive returns to students' initiative in class, university aspirations and expectations, as well as parental money investment in the form of private tutoring, parental support and university aspirations for their child. There is evidence of negative returns to students' academic self-efficacy, and parental strictness and harshness. The differences between total and partial average returns reflect the fact that many of these inputs are correlated.

Figure 6 puts together the results from Figures 4 and 5 to produce estimates of the mediated effect of peer effects by our measured educational inputs, as per equation (2). The bar in green shows that our mechanisms explain a negative and statistically but not economically significant amount of our estimated peer effect—which means that the effect of higher-achieving peers on these inputs and their estimated return jointly make it *harder*, not easier, to explain the academic peer effects. Jointly, all our educational inputs explain only -0.9 percent of a standard deviation of the 5.4 percent of a standard deviation academic peer effect. This negative mediation is chiefly driven by the negative effects of higher-achieving peers on effort combined with the large and positive estimate of the returns to effort on academic achievement. None of the other inputs we consider has a statistically or economically significant mediating effect.

Figure 6. Academic peer effects mediated by educational inputs



This figure reports the mediated effects based on Gelbach's (2016) decomposition of our academic peer effect estimate using only within-school variation in our estimation sample containing 227 schools, 853 classes, and 12,816 students. These estimates are produced using a modified version of the `b1x2` *Stata package. Rows present the mediated effect of different educational inputs in wave 2. All models control for school fixed effects, student test scores, average peer test scores, and educational inputs in wave 1. The total mediated effect is shown in green, and student, parent, and school & teacher inputs are shown in navy blue, maroon, and teal. p-values shown are based on standard errors clustered at the classroom level. These results are also available in Appendix Table B4.*

Overall, the results in this section show that, in spite of having precise estimates of *i*) academic peer effects and of *ii*) the effects of higher-achieving peers on educational inputs, which could potentially act as mechanisms for these peer effects, our potential mechanisms explain practically nothing of peer effects. These new results show the difficulties of learning about the mechanisms that drive social interaction effects and suggest that the prevailing microeconometric approach to exploring these mechanisms can be of limited use. Puzzling results such as these open a number of questions and can prove to be a knowledge base to build on, as long as its foundations are solid. Precisely because of this, in the next section we show that our main results and conclusions are robust to a myriad of specification checks and potential concerns. In particular, section 5.4. shows that these results on the absence of mediation are not hiding heterogeneity either in the sense that higher-achieving peers affect subgroups of students differently, or in the sense that subgroups of students are affected by higher-achieving peers through different mechanisms.

5 Sensitivity Analyses

In this section, we discuss the sensitivity of our results along four dimensions: *i*) robustness to changes in our identification strategy; *ii*) robustness to the effects of measurement error in our data; *iii*) robustness of our inference to different constructions of standard errors; and i*v*) robustness of our conclusions on the mediation analyses to the presence of heterogeneous peer effects.

5.1 Robustness of our identification strategy

Here, we first provide additional evidence of random assignment of students to classrooms within schools in our trimmed sample using permutation-based sorting tests, and using non-parametric sorting tests. Many of these tests have become standard in the empirical peer effects literature. We then exploit the richness of our data—in particular the fact that we observe many pre-assignment characteristics of students, parents and teachers—to show that proportional selection on unobservable characteristics is very unlikely to be driving our results.

5.1.1 Permutation-based sorting tests

In the empirical peer effects literature, permutation-based tests of random assignment of students to peer groups have become very popular. These tests compared the actual student group

composition in the data to counterfactual compositions simulated under the null of random assignment, as described in Section 3.2. As an additional check for random assignment in our data, we estimate permutation-based sorting tests akin to those in e.g., Carrell and West (2010) and Lim and Meer (2017, 2019) in our trimmed sample.

For these tests, we simulate 10,000 classrooms under the null of random assignment of students to classrooms within schools. We do so by randomly drawing sampled students with replacement and keeping the core structure of the data—respecting students' assignment to schools, and number and size of classrooms within each school. We then calculate the mean of our key pre-treatment characteristics in each of the 10,000 synthetic classrooms. Finally, for each classroom, we count the times the synthetic classroom mean of each characteristic was more extreme than the actual classroom mean, relative to the schools mean. The share of times this happens corresponds to the classroom-level empirical p-value of a test of random assignment of students to classrooms within schools based on that characteristic.

Appendix Table B5 shows these permutation-based empirical p-values for each key pre-determined characteristic separately. Under random assignment, the shares in the second through fourth column should be close to the nominal rejection rates of 0.10, 0.05 and 0.01 in most or all rows. The evidence in this table strongly supports the idea of random assignment to classrooms within schools in our trimmed sample.

5.1.2 Non-parametric sorting test

As implemented, balancing tests and sorting tests all have one important shortcoming: their linearity. Balancing tests, for example, assess whether female students are assigned to higher-achieving peers. Sorting tests try to capture whether female students end up in classrooms with other female students. But these tests do not truly test for what random assignment would imply: whether classrooms systematically differ in these pre-assignment characteristics *in any way*. In other words, these tests do not test non-parametrically for systematic assignment of students to classrooms. A few studies do use this non-parametric sorting test (Ammermueller and Pischke, 2009; Sojourner, 2013; Feld and Zölitz, 2017).

We implement this test in the following steps. First, we estimate school-by-school regressions of each pre-assignment characteristic on a set of classroom dummies. Second, we jointly test the statistical significance of these classroom dummies and collect the p-values of these tests. We end up with a set of 2,790 p-values; one for each of the 227 schools in our sample and each of our key 18 pre-assignment characteristics. We then note that, under the null of random assignment of classrooms to schools, these p-values should be uniformly distributed. Therefore, as a third step, we check whether more than ten, five and one percent of the school-level p-values fall under the nominal values 0.10, 0.05 and 0.01 for each characteristic.

Appendix Table B6 shows empirical p-value distributions for each characteristic separately. Consistent with our tests in Section 3.4, these results also show some evidence of minor sorting based on intellectual curiosity, gifted arts classroom enrolment, and parents pushing for assignment to particular classrooms. Overall, however, these tests provide yet again evidence in strong support of random assignment to classrooms within schools in our trimmed sample.

5.1.3 Proportional selection on unobservable to observable characteristics

Our trimmed sample is chosen in a data-driven way that ensures that key pre-assignment characteristics are unrelated to average peer test scores. This identification strategy relies on our ability to find data that reflects a clean quasi-experiment in classroom allocation, yet systematically excludes entire schools from our sample, which might lead to sample selection issues. Still, we ask ourselves whether the few observable characteristics that remain correlated to higher-ability peers could present reasonable concerns about unobserved heterogeneity. This calls for an analysis of proportional selection on observable characteristics, as discussed in Altonji et al. (2005) and Oster (2019). The two conditions for this analysis to make sense are *i*) that our observable characteristics for these analyses are a random sample of all determinants of student achievement, and *ii*) that the number of observed and unobserved determinants of student achievement are large and neither element is dominating. Along the argument lines of Altonji et al. (2005), we assume that the TEPS fulfils both conditions.

We implement this analysis by calculating Oster's (2019) δ, the share of proportional selection needed to explain away the entire peer effect we estimate. Values of $\delta > 1$ imply that the selection on unobservable characteristics would need to be at least as large as the selection on observable

characteristics to explain away the entire peer effect estimate, which, given the data and data context, is an unreasonable assumption. A $\delta < 1$ implies that the omitted variable bias from unobservable variables positively correlated with the observable variables included would bias the peer effects *away from zero*, not towards, and should therefore not be concerning as confounders. In this type of analysis, thus, finding values of δ between zero and one is worrisome, and could indicate a potential concern for unobserved selection affecting results. The observables we use for these analyses are extensive: they include our balancing controls (household income, family engagement with homework, gifted art classroom assignment, and parents' pushiness to get child assigned to a particular classroom) and our standardized measures of student, parent, school and teacher educational inputs in wave 1. Assuming that selection on unobservable characteristics occurs in proportion to the selection on this set of variables implies, by exclusion, that school fixed effects and students' own test scores in wave 1—*a priori* essential for our identification strategy and standard in the literature—cannot inform the proportional selection analyses. We also use a hypothetical maximum R-Squared value of 1.3 times the R-Squared of the unrestricted model, which is the standard choice for these analyses.

Appendix Table B7 shows Oster's δ for all our main estimates estimating using Stata's `psacalc` command. For nearly all our estimates, Oster's δ is negative which implies that proportionally selection on unobserved confounders are unlikely to explain our effects. The one exception is the δ of 0.10 for the effect of higher-achieving peers on parental investments in tutoring, which is anyway insignificantly different from zero so none of our conclusions change following the results of this analysis. Overall, we conclude that proportional selection on unobservable variables cannot explain away any of our findings.

5.2 Robustness to measurement error and selective attrition

We now turn our focus to the measurement error in our data We show that our main estimates *i*) are robust to using different measures of student and peer academic ability, *ii*) are not attenuated by measurement error in average peer test scores, and *iii*) are not biased by the fact that we do not observe whole classrooms.

5.2.1 Main results with alternative measures of ability

Our main results use the TEPS scores in the comprehensive ability test. As discussed in Section 3.2, this test was designed by TEPS team and uses 75 multiple-choice question to measure of students' cognitive ability and analytical reasoning. However, after a series of factor analyses and after estimating 3-parameter Item Response Theory (IRT) models, the TEPS team could also identify two highly correlated but distinct subcomponents measuring analytical ability and mathematical ability based on disjoint subsets of test questions. The IRT models were also used to produce the standardized Bayesian posterior means of the three components identifiable in the test—the general ability component and the analytical ability and mathematical ability subcomponents.[7]

Appendix Table B8 shows that our main results are robust to using the analytical and mathematical subcomponents of the comprehensive ability test scores as measures of student and peer ability (columns 1 and 2). Our main results are also robust to using the Bayesian posterior means of these components, arguably a more precise and efficient measure of ability (columns 3 through 5).

5.2.2 Correction for classical measurement error in peer ability

Even in excellent measures of student and peer ability, such as the well-designed standardized test scores in TEPS, there will still be some measurement error. Under random assignment and with classical measurement error (i.e., independent of all covariates and of true ability), this measurement error will attenuate our peer effect estimates (Sojourner, 2013; Feld and Zölitz, 2017). We can address this attenuation bias in two similar ways. Noting that the analytical and mathematical subcomponents of test scores are measured with disjoint sets of questions, we can use average peer test scores using one subcomponent as an instrument for average peer test scores using the other in an instrumental variable (IV) estimator. See e.g., Salamanca et al. (2020) for a similar approach to account for measurement error in personality traits. This approach would eliminate attenuation bias from classical measurement error under two assumption: *i*) that both subcomponents have a strong common element of overall ability, and *ii*) that measurement error

[7] See http://www.teps.sinica.edu.tw/description/TestingReport2004-2-10.pdf (in Mandarin) for a description of these analyses.

in test questions is uncorrelated across subcomponents. The first assumption is well supported by our data and by the TEPS team factor and IRT analyses. The second assumption is stronger; if it does not hold it would result in some attenuation bias left in the IV estimate.

Appendix Table B9 shows that, although less precisely estimated due to the usual efficiency loss from instrumental variable models, the IV point estimates are near-identical to our main results (columns 1 and 2). We thus view this as evidence of little attenuation bias due to classical measurement error in our estimates.

One potential problem with the estimators above is that the IV estimates need to be interpreted as academic peer effects in *analytical* and *mathematical* ability, rather than in *comprehensive* ability. We address this problem by constructing a 'mixed IV' estimator. In this estimator, we first construct an ability measure that, for each student, is randomly defined as either the analytical subcomponent score or the mathematical subcomponent score with equal probability. This ability measure is therefore an equal-weighted average of the analytical and mathematical subcomponents and can be interpreted as measuring general ability. We call this our 'mixed ability' measure. We also construct an ability instrument that is defined by the same random process to be *the subcomponent that was not assigned as ability*. For example, if for student i ability is measured as the analytical subcomponent score, then the ability instrument is defined as the mathematical subcomponent score. We call this our 'mixed ability instrument'. Under the same assumptions above, an IV estimate that instruments our mixed ability with our mixed ability instrument also corrects for attenuation bias while identifying academic peer effects using general ability, rather than analytical or mathematical ability. We show that this new estimator produces very similar results to our main peer effect on test scores (Appendix Table B9, column 3). It also produces slightly larger estimated magnitudes of the effect of higher-achieving peers on study effort and students' university aspiration and expectations, and similar estimates for the effect on parental time investments (Appendix Table B10). Back-of-the-envelope calculations show that these slightly larger estimates do not change our conclusions on the mediated effects of higher-achieving peers. We thus conclude that measurement error does not alter any of our main findings.

5.2.3 Sojourner (2013) correction for incomplete classroom sampling

Many empirical peer effect studies, including ours, has incomplete classroom data which results in incomplete sampling of students' peer group. Sojourner (2013) shows that this issue can result in bias in peer effect estimates that is similar to classical attenuation bias under random assignment, and much more difficult to sign and quantify under non-random assignment. He also proposes a correction for this bias that relies on *i*) weighting estimates by the share of peers sampled and *ii*) controlling for these shares at the school level. Often these last controls are multicollinear with the weighted peer measures, so he also suggests less restrictive estimators that control for the share of peers sampled within predetermined school clusters. We implement both methods in our data to evaluate the extent of this bias in our main results. The left-most column on the table implements Sojourner's preferred correction which can lead to substantial loss of power because it heavily restricts the identifying variation used by the estimator. The second through sixth columns implement specifications which trade off more power for less bias reduction, from left to right.

Appendix Table B11 shows substantially larger effects of higher-achieving peers on student test scores and proportionally larger effects on students' university aspirations and expectations and parental time investments. This is all consistent with Sojourner's findings and with the data originating from conditional random assignment to classrooms within schools. The analyses do not reveal other effects of higher-ability peers. Moreover, since the attenuation of all our estimated effects is proportional, our conclusions about mediated peer effects remains unchanged. This suggests that not observing complete classrooms in our data could lead to understating the importance of academic peer effects, but does not affect our (in)ability to explain their mechanisms.

5.3 Randomization inference and multiple hypotheses testing correction

Having established the robustness of our point estimates of peer effects, in this subsection we show that our inference on these effects is robust to *i*) constructing standard errors based on recent randomization inference techniques and *ii*) to accounting for multiple hypotheses testing in our standard error calculations.

We first reassess inference on our main results using Young's (2019) randomization-*t* procedure. Our analyses benefit from this procedure because of the potential influence of a few high-leverage students, classrooms or schools, and we want to ensure that our inference is robust to this occurrence. We also want to use inference that does not make strong assumption on the structure of error terms given the complexity of the TEPS sampling design and peer treatment. Other benefits of randomization inference, such as *i*) correcting for few treatment clusters or *ii*) issues of joint testing are less important for this study, because *i*) we observe several classrooms per school, and ii) each regression has one treatment effect of interest.

We construct randomization-*t* based empirical p-values via a very similar simulation procedure to the one used for permutation tests. The key difference is that, in each simulation, we capture the *t*-statistics of interest—the coefficient of the key variable of interest divided by its cluster-robust standard error—and construct empirical p-values based the share of occurrences where simulated *t*-statistics are more extreme than our actual *t*-statistic of interest. We use 10,000 simulations of random assignment to classroom within schools to produce randomization-*t* empirical p-values for our main results. Appendix Table B12 shows that when using randomization-*t* inference p-values for conducting inference, our main conclusions on the effects of higher-achieving peers hold at the 5% significance level for student achievement and parental time investments, and at the 10% significance level for student university aspirations and expectations.

In a second analysis, we adjust our inference for multiple hypotheses testing: the problem that the chance of falsely rejecting a correct null hypothesis increases with the number of tests performed. We adjust for this by implementing the Romano-Wolf multiple hypothesis correction (Romano and Wolf, 2005a,b) using Stata's `rwolf` command (Clarke et al., 2019). This procedure ensures that the familywise error rate—the probability of committing at least one Type I error across a set of hypotheses tested—does not exceed its predetermined significance. We consider all our main results to be part of the same family of tests. Appendix Table B12 shows that our main conclusions on the effect of higher-achieving peers on student achievement and on parental time investment hold at the 10% significance level, but our evidence on students' university aspirations and expectations now appear not to be statistically significant.

Overall, with these different inference methods we still find strong evidence of academic peer effects in our data but somewhat weaker evidence of significant effects on educational inputs. This reinforces our conclusions of no mediated effects for academic peer effects.

5.4 Heterogeneous peer effects and robustness of mediated effects

Finally, we explore the sensitivity of our mediation analyses. Our chief concern here is the possibility that our lack of meaningful mediation can occur not because educational inputs cannot explain academic peer effects, but rather as the result of heterogeneity peer effects across subgroups. Heterogeneity can occur in two forms: firstly, academic peer effects could vary widely across subgroups—a result found in several studies across ability (Carrell, Fullerton and West, 2009), gender (Whitmore, 2005; Lavy and Schlosser, 2011), race (Hoxby, 2000, Hoxby and Weingarth, 2005), but secondly and perhaps most importantly, the drivers of peer effects for each subgroup could also widely differ, as suggested by Brady, Insler and Rahmam (2017). For example, higher-achieving peers could improve test scores of low-ability students because they reduce the amount of classroom disruption (see e.g. Lavy, Paserman and Schlosser, 2011) and improve test scores of high-ability students because they increase effort. Yet we might be unable to detect enough mediation via truancy and effort on the *average* academic peer effect. This form of heterogeneity would wrongly lead us to conclude that truancy and effort cannot explain at least part of academic peer effects. One way to assess whether this particular type of heterogeneity is a likely explanation for our findings is to estimate the heterogeneity of peer effects *and* their mediation via educational inputs across various subgroups.

There are countless dimensions to explore heterogeneity in academic peer effects in our data. Based on existing heterogeneous effects in the academic peer literature, and on a broader literature on the sociodemographic predictors of student test scores, we chose to explore peer effect and mediation heterogeneity across: student ability, gender, household income, parental education, public vs private schooling, and teacher experience. Appendix Table B13 shows that, by and large, there is little subgroup heterogeneity in our estimated academic peer effects and their mediation. Academic peer effects are slightly larger at the top and middle of the student ability distribution and with highly experienced teachers, yet are the same across student gender, household income, parental education. More importantly, our inputs can still mediate either small or negative parts of

these academic peer effect for any one of these subgroups. Altogether, we show strong evidence of little heterogeneity in academic peer effects and in mediated effects.

Based on these results we conclude that subgroup heterogeneity is not a likely explanation for the fact that our many educational inputs do not mediate academic peer effects.

6 Conclusions

We estimate the effect of being randomly assigned to classrooms with higher-achieving peers on students' standardized test scores two years later, and on many other intermediate outcomes of students, their parents, and their teachers. We conduct a formal mediation analysis of academic peer effects to explore several potential mechanisms, one of the first ones of its kind in a field with over twenty years of research and hundreds of articles. Our study thus gives the most comprehensive view of how much academic peer effects are explained by changes in educational investments in a setting with a credible identification strategy.

Students assigned to classroom peers with one standard deviation higher test scores at the beginning of middle school experience an economically sizable 5.4 percent of a standard deviation increase in their own standardized test scores two years later. These higher-achieving classroom peers also *decrease* student school effort, increase students' aspirations and expectations to go to university, and increase the time parents spend with them. Higher-achieving peers have precisely estimated zero effects on many other educational inputs we consider, including initiative in class, mental health, cheating on exams or truant behavior.

For producing these results, we use data in a setting with a well-documented country-wide mandate of random assignment of students to classrooms within schools. The data, however, shows that this random assignment was likely not upheld everywhere, which is not entirely surprising: we can think of legal and illegal ways in which sorting can still occur—for example, via allowed "talent" classrooms in schools, or due to principals sorting students into classrooms in defiance of the mandate. Similar violations to national mandates are common in similar settings (e.g., Gong, Lu, & Song, 2019; Eble & Hu, 2019). We develop a data-driven procedure to remove schools likely to be defying the mandate of random assignment from our estimation sample and show that data in this trimmed sample is strongly consistent with random assignment. This fishing algorithm can

be used to improve quasi-experimental designs in settings where random assignment to peers is suspected to be violated in some, but not all, assignment groups. It can more generally be used in any setting where imperfect compliance of (quasi-)experimental treatment assignment is suspected.

Our findings contribute to diverse literatures on the effects of higher-achieving peers on various educational inputs (e.g., Bursztyn, and Jensen, 2015; Booij, Leuven, and Oosterbeek, 2017; Bursztyn, Egorov, and Jensen, 2018), the determinants of educational aspirations (e.g., La Ferrara, 2018; Carlana, 2019), and parental reactions to school inputs (e.g., Pop-Eleches & Urquiola, 2013; Fredriksson et al. 2013; Cobb-Clark et al. 2020).

Yet we also find that either of these intermediate effects, nor all of them as a whole, are able to account for the large increase in scores that students experience down the road.

Since our academic peer effects remain largely unexplained, even after covering such a large battery of agents and inputs in the education production function, it could be tempting to conclude that academic peer effects are unexplainable by current methods. It could also be tempting to resort to new ways of modeling social interactions, rather than provide more empirical findings using tried and true data and empirical designs. Instead, we prefer to go back to Manski's (2000) seminal work, in which he explained how crucial standard economic concepts are for modeling social interactions: expectations, preferences, market incentives. Our results show evidence for both core types of social interactions described by Manski — peer effects in expectations and peer effects in outcomes. We therefore see our results an invitation to study how peer effects in expectations translate into choices performance (or why they fail to do so). We are also eager to explore how these processes might systematically vary across settings and circumstances, including the role of competitive environments, the endogenous choice of peers within classrooms, whether peer interactions are invigilated and controlled, and by whom. Recent work by e.g. Burstzyn and Jensen (2015), Bedard and Fischer (2019), and Babcock et al. (2019) is already making great progress in this direction.

Our results also get us closer to using peer effects to confidently inform and design classroom assignment policies. A pervasive concern with systematic assignment policies is that their benefits might come with unmeasured cost on, e.g. classroom disruption, increasing stress, deteriorating

mental health, and higher effort to keep up with one's higher-achieving peers. Our study shows that many of these concerns are unfounded. And yet, like other studies, we find sufficiently non-linear peer effects across student ability to make Pareto-improving class assignment policies a possibility (See Appendix Table B13). Designing such class reallocation improvements can be difficult, and even then the expected gains might not be realized (Graham, Imbens & Ridder, 2014; Carrell et al., 2013). Nevertheless, our findings rule out many behavioral reactions to higher-achieving peers that could complicate the design of such policies, and further rule out several usually unobserved costs that could subtract from any realized gains. Altogether, this brings back the possibility of class reassignment policies that improve student welfare.

References

Abdulkadiroğlu, A., Angrist, J. and Pathak, P., 2014. The elite illusion: Achievement effects at Boston and New York exam schools. *Econometrica*, *82*(1), pp.137-196.

Agostinelli, F., 2018. Investing in children's skills: An equilibrium analysis of social interactions and parental investments. *Unpublished Manuscript, Arizona State University*.

Agostinelli, F., Doepke, M., Sorrenti, G. and Zilibotti, F., 2020. *It takes a village: the economics of parenting with neighborhood and peer effects* (No. w27050). National Bureau of Economic Research.

Ahern, K.R., Duchin, R. and Shumway, T., 2014. Peer effects in risk aversion and trust. *The Review of Financial Studies*, *27*(11), pp.3213-3240.

Altonji, J.G., Elder, T.E. and Taber, C.R., 2005. Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy*, 113(1), pp.151-184.

Ammermueller, A. and Pischke, J.S., 2009. Peer effects in European primary schools: Evidence from the progress in international reading literacy study. *Journal of Labor Economics*, *27*(3), pp.315-348.

Angrist, J.D., 2014. The perils of peer effects. *Labour Economics*, 30, pp.98-108.

Angrist, J.D. and Lang, K., 2004. Does school integration generate peer effects? Evidence from Boston's Metco Program. *American Economic Review*, *94*(5), pp.1613-1634.

Arcidiacono, P. and Nicholson, S., 2005. Peer effects in medical school. *Journal of public Economics*, *89*(2-3), pp.327-350.

Argys, L.M. and Rees, D.I., 2008. Searching for peer group effects: A test of the contagion hypothesis. *The Review of Economics and Statistics*, *90*(3), pp.442-458.

Aucejo, E.M., Coate, P., Fruehwirth, J., Kelly, S. and Mozenter, Z., 2018. Teacher effectiveness and classroom composition. Unpublished manuscript.

Azmat, G., Bagues, M., Cabrales, A. and Iriberri, N., 2019. What You Don't Know… Can't Hurt You? A Natural Field Experiment on Relative Performance Feedback in Higher Education. *Management Science*, *65*(8), pp.3714-3736.

Babcock, P., Bedard, K., Fischer, S. and Hartman, J., 2019. *Coordination and Contagion: Individual Connections and Peer Mechanisms in a Randomized Field Experiment* (No. 1904).

Balsa, A., Gandelman, N. and Roldán, F., 2018. Peer and parental influence in academic performance and alcohol use. *Labour Economics*, *55*, pp.41-55.

Beaman, L., Duflo, E., Pande, R. and Topalova, P., 2012. Female leadership raises aspirations and educational attainment for girls: A policy experiment in India. *Science*, 335(6068), pp.582-586.

Bedard, K. and Fischer, S., 2019. Does the response to competition depend on perceived ability? Evidence from a classroom experiment. *Journal of Economic Behavior & Organization*, *159*, pp.146-166.

Benjamini, Y. and Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), pp.289-300.

BenYishay, A. and Mobarak, A.M., 2018. Social learning and incentives for experimentation and communication. *The Review of Economic Studies*, *86*(3), pp.976-1009.

Bietenbeck, J., 2015. The Long-Term Impacts of Low-Achieving Childhood Peers: Evidence from Project STAR. *Journal of the European Economic Association*.

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

Bifulco, R., Fletcher, J.M., Oh, S.J. and Ross, S.L., 2014. Do high school peers have persistent effects on college attainment and other life outcomes?. *Labour economics*, *29*, pp.83-90.

Black, S.E., Devereux, P.J. and Salvanes, K.G., 2013. Under pressure? The effect of peers on outcomes of young adults. *Journal of Labor Economics*, *31*(1), pp.119-153.

Boisjoly, J., Duncan, G.J., Kremer, M., Levy, D.M. and Eccles, J., 2006. Empathy or antipathy? The impact of diversity. *American Economic Review*, *96*(5), pp.1890-1905.

Bonesrønning, H., 2004. The determinants of parental effort in education production: do parents respond to changes in class size? *Economics of Education Review*, *23*(1), pp.1-9.

Bonilla-Mejía, L., Bottan, N.L. and Ham, A., 2019. Information policies and higher education choices experimental evidence from Colombia. *Journal of Behavioral and Experimental Economics*, 83, p.101468.

Boozer, M. and Cacciola, S.E., 2001. Inside the 'Black Box' of Project STAR: Estimation of peer effects using experimental data. *Yale Economic Growth Center Discussion Paper*, (832).

Brady, R.R., Insler, M.A. and Rahman, A.S., 2017. Bad Company: Understanding negative peer effects in college achievement. *European Economic Review*, *98*, pp.144-168.

Brenøe, A. and Zölitz, U. Exposure to more female peers widens the gender gap in STEM participation. *Journal of Labor Economics, 2019*

Brunello, G., De Paola, M. and Scoppa, V., 2010. Peer effects in higher education: Does the field of study matter? *Economic Inquiry*, *48*(3), pp.621-634.

Burke, M.A. and Sass, T.R., 2013. Classroom peer effects and student achievement. *Journal of Labor Economics*, *31*(1), pp.51-82.

Bursztyn, L. and Jensen, R., 2015. How does peer pressure affect educational investments? *The quarterly journal of economics*, *130*(3), pp.1329-1367.

Bursztyn, L., Ederer, F., Ferman, B. and Yuchtman, N., 2014. Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions. *Econometrica*, *82*(4), pp.1273-1301.

Bursztyn, L., Egorov, G. and Jensen, R., 2018. Cool to be smart or smart to be cool? Understanding peer pressure in education. *The Review of Economic Studies*, *86*(4), pp.1487-1526.

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

Card, D. and Giuliano, L., 2013. Peer effects and multiple equilibria in the risky behavior of friends. *Review of Economics and Statistics*, *95*(4), pp.1130-1149.

Carlana, M., 2019. Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, *134*(3), pp.1163-1224.

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

Carrell, S.E., Hoekstra, M. and Kuka, E., 2018. The long-run effects of disruptive peers. *American Economic Review*, *108*(11), pp.3377-3415.

Carrell, S.E., Malmstrom, F.V. and West, J.E., 2008. Peer effects in academic cheating. *Journal of human resources*, *43*(1), pp.173-207.

Carrell, S.E., Sacerdote, B.I. and West, J.E., 2013. From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica*, *81*(3), pp.855-882.
Carrell, S.E. and West, J.E., 2010. Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, *118*(3), pp.409-432.
Chang, S., Cobb-Clark, D. A., and Salamanca, N., 2020. Parents' Responses to Teacher Qualifications. *IZA Discussion Paper Series, No. 13065*
Chetty, R., Friedman, J.N., Hilger, N., Saez, E., Schanzenbach, D.W. and Yagan, D., 2011. How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly journal of economics*, *126*(4), pp.1593-1660.
Clark, D., 2010. Selective schools and academic achievement. *The BE Journal of Economic Analysis & Policy*, *10*(1).
Clarke, D., Romano, J.P. and Wolf, M., 2019. The Romano-Wolf Multiple Hypothesis Correction in Stata. *IZA Discussion Paper Series, No. 12845*
Correia, S., 2018. REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects. Statistical Software Components S457874, *Boston College Department of Economics*, revised 18 Nov 2019.
Cunha, F. and Heckman, J., 2007. The technology of skill formation. *American Economic Review*, 97(2), pp.31-47.
Datar, A. and Mason, B., 2008. Do reductions in class size "crowd out" parental investment in education?. *Economics of Education Review*, *27*(6), pp.712-723.
Dee, T.S., 2004. Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, *86*(1), pp.195-210.
Deming, D.J., 2011. Better schools, less crime?. *The Quarterly Journal of Economics*, *126*(4), pp.2063-2115.
De Paola, M. and Scoppa, V., 2010. Peer group effects on the academic performance of Italian students. *Applied Economics*, *42*(17), pp.2203-2215.
Diette, T.M. and Uwaifo Oyelere, R., 2014. Gender and race heterogeneity: The impact of students with limited english on native students' performance. *American Economic Review*, *104*(5), pp.412-17.
Dobbie, W. and Fryer Jr, R.G., 2014. The impact of attending a school with higher-achieving peers: Evidence from the New York City exam schools. *American Economic Journal: Applied Economics*, *6*(3), pp.58-75.
Duflo, E., Dupas, P. and Kremer, M., 2011. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, *101*(5), pp.1739-74.
Eble, A. and Hu, F., 2019. How important are beliefs about gender differences in math ability? Transmission across generations and impacts on child outcomes. *CDEP-CGEG Working Paper*, *53*.
Elsner, B. and Isphording, I.E., 2017. A big fish in a small pond: Ability rank and human capital investment. *Journal of Labor Economics*, *35*(3), pp.787-828.
Feld, J. and Zölitz, U., 2017. Understanding peer effects: On the nature, estimation, and channels of peer effects. *Journal of Labor Economics*, *35*(2), pp.387-428.
Feng, H. and Li, J., 2016. Head teachers, peer effects, and student achievement. *China Economic Review*, *41*, pp.268-283.
Figlio, D.N., 2007. Boys named Sue: Disruptive children and their peers. *Education finance and policy*, *2*(4), pp.376-394.

Figlio, D. and Özek, U., 2019. Unwelcome guests? The effects of refugees on the educational outcomes of incumbent students. *Journal of Labor Economics*, *37*(4), pp.1061-1096.

Finn, J.D., Pannozzo, G.M. and Achilles, C.M., 2003. The "why's" of class size: Student behavior in small classes. *Review of Educational Research*, *73*(3), pp.321-368.

Fiorini, M. and Keane, M.P., 2014. How the allocation of children's time affects cognitive and noncognitive development. *Journal of Labor Economics*, 32(4), pp.787-836.

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

Fredriksson, P., Öckert, B. and Oosterbeek, H., 2016. Parental responses to public investments in children: Evidence from a maximum class size rule. *Journal of Human Resources*, *51*(4), pp.832-868.

Fruehwirth, J.C., 2014. Can achievement peer effect estimates inform policy? A view from inside the black box. *Review of Economics and Statistics*, *96*(3), pp.514-523.

Garlick, R., 2018. Academic Peer Effects with Different Group Assignment Policies: Residential Tracking versus Random Assignment. *American Economic Journal: Applied Economics*, *10*(3), pp.345-69.

Gelbach, J.B., 2016. When do covariates matter? And which ones, and how much?. *Journal of Labor Economics*, *34*(2), pp.509-543.

Gibbons, S. and Telhaj, S., 2016. Peer effects: Evidence from secondary school transition in England. *Oxford Bulletin of Economics and Statistics*, *78*(4), pp.548-575.

Gillen, B., Snowberg, E. and Yariv, L., 2019. Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, *127*(4), pp.1826-1863.

Golsteyn, B., Non, A. and Zölitz, U., forthcoming. The impact of peer personality on academic achievement. *Journal of Political Economy*

Gong, J., Lu, Y. and Song, H., 2019. Gender peer effects on students' academic and noncognitive outcomes: Evidence and mechanisms. *Journal of Human Resources*, pp.0918-9736R2.

Gould, E.D., Lavy, V. and Paserman, M.D., 2004. Immigrating to opportunity: Estimating the effect of school quality using a natural experiment on Ethiopians in Israel. *The Quarterly Journal of Economics*, *119*(2), pp.489-526.

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

Graham, B.S., Imbens, G.W. and Ridder, G., 2014. Complementarity and aggregate implications of assortative matching: A nonparametric analysis. *Quantitative Economics*, 5(1), pp.29-66.

Griffith, A.L. and Rask, K.N., 2014. Peer effects in higher education: A look at heterogeneous impacts. *Economics of Education Review*, *39*, pp.65-77.

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

Hanushek, E.A., Kain, J.F. and Rivkin, S.G., 2009. New evidence about Brown v. Board of Education: The complex effects of school racial composition on achievement. *Journal of labor economics*, *27*(3), pp.349-383.

Hanushek, E.A., Kain, J.F., Markman, J.M. and Rivkin, S.G., 2003. Does peer ability affect student achievement? *Journal of applied econometrics*, *18*(5), pp.527-544.

Hoekstra, M., 2009. The effect of attending the flagship state university on earnings: A discontinuity-based approach. *The Review of Economics and Statistics*, *91*(4), pp.717-724.

Hoekstra, M., Mouganie, P. and Wang, Y., 2018. Peer quality and the academic benefits to attending better schools. *Journal of Labor Economics*, *36*(4), pp.841-884.

Hong, S.C. and Lee, J., 2017. Who is sitting next to you? Peer effects inside the classroom. *Quantitative Economics*, 8(1), pp.239-275. Vancouver

Hoxby, C., 2000. *Peer effects in the classroom: Learning from gender and race variation* (No. w7867). National Bureau of Economic Research.

Hoxby, C.M. and Weingarth, G., 2005. *Taking race out of the equation: School reassignment and the structure of peer effects* (No. 7867). Working paper.

Huntington-Klein, N. and Rose, E., 2018, May. Gender peer effects in a predominantly male environment: Evidence from West point. In *AEA Papers and Proceedings* (Vol. 108, pp. 392-95).

Imai, K., Tingley, D. and Yamamoto, T., 2013. Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A* (Statistics in Society), 176(1), pp.5-51.

Imberman, S.A., Kugler, A.D. and Sacerdote, B.I., 2012. Katrina's children: Evidence on the structure of peer effects from hurricane evacuees. *American Economic Review*, *102*(5), pp.2048-82.

Jackson, C.K., 2013. Can higher-achieving peers explain the benefits to attending selective schools? Evidence from Trinidad and Tobago. *Journal of Public Economics*, *108*, pp.63-77.

Jacob, B.A. and Lefgren, L., 2007. What do parents value in education? An empirical investigation of parents revealed preferences for teachers. *The Quarterly Journal of Economics*, *122*(4), pp.1603-1637.

Jain, T. and Kapoor, M., 2015. The impact of study groups and roommates on academic performance. *Review of Economics and Statistics*, *97*(1), pp.44-54.

Kang, C., 2007. Classroom peer effects and academic achievement: Quasi-randomization evidence from South Korea. *Journal of Urban Economics*, *61*(3), pp.458-495.

Karbownik, K., 2020. The effects of student composition on teacher turnover: Evidence from an admission reform. *Economics of Education Review (*forthcoming)

Keele, L., Tingley, D. and Yamamoto, T., 2015. Identifying mechanisms behind policy interventions via causal mediation analysis. *Journal of Policy Analysis and Management*, 34(4), pp.937-963.

Kiss, D., 2013. The impact of peer achievement and peer heterogeneity on own achievement growth: Evidence from school transitions. *Economics of Education Review*, *37*, pp.58-65.

Kramarz, F., Machin, S. and Ouazad, A., 2015. Using compulsory mobility to identify school quality and peer effects. *Oxford Bulletin of Economics and Statistics*, *77*(4), pp.566-587.

Kremer, M. and Levy, D., 2008. Peer effects and alcohol use among college students. *Journal of Economic perspectives*, *22*(3), pp.189-206.

Krueger, A.B., 1999. Experimental estimates of education production functions. *The quarterly journal of economics*, *114*(2), pp.497-532.

Krueger, A.B. and Whitmore, D.M., 2001. The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *The Economic Journal*, *111*(468), pp.1-28.

La Ferrara, E., 2019. Aspirations, Social Norms, and Development. *Journal of the European Economic Association*, *17*(6), pp.1687-1722.
Lavy, V. and Schlosser, A., 2011. Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics*, *3*(2), pp.1-33.
Lavy, V., Paserman, M.D. and Schlosser, A., 2011. Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom. *The Economic Journal*, *122*(559), pp.208-237.
Lavy, V., Silva, O. and Weinhardt, F., 2012. The good, the bad, and the average: Evidence on ability peer effects in schools. *Journal of Labor Economics*, *30*(2), pp.367-414.
Law, W. W. (2004). Translating globalization and democratization into local policy: Educational reform in Hong Kong and Taiwan. *International Review of Education*, 50, 497-524.
Lazear, E.P., 2001. Educational production. *The Quarterly Journal of Economics*, *116*(3), pp.777-803.
Lim, J. and Meer, J., 2017. The impact of teacher–student gender matches random assignment evidence from South Korea. *Journal of Human Resources*, *52*(4), pp.979-997.
Lim, J. and Meer, J., 2019. Persistent effects of teacher-student gender matches. *Journal of Human Resources*, pp.0218-9314R4.
Lu, F. and Anderson, M.L., 2014. Peer effects in microenvironments: The benefits of homogeneous classroom groups. *Journal of Labor Economics*, *33*(1), pp.91-122.
Lyle, D.S., 2007. Estimating and interpreting peer and role model effects from randomly assigned social groups at West Point. *The Review of Economics and Statistics*, *89*(2), pp.289-299.
Manski, C.F., 1993. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, *60*(3), pp.531-542.
Manski, C.F., 2000. Economic analysis of social interactions. *Journal of economic perspectives*, *14*(3), pp.115-136.
Marmaros, D. and Sacerdote, B., 2002. Peer and social networks in job search. *European economic review*, *46*(4-5), pp.870-879.
McEwan, P.J., 2003. Peer effects on student achievement: Evidence from Chile. *Economics of education review*, *22*(2), pp.131-141.
Moretti, E., 2011. Social learning and peer effects in consumption: Evidence from movie sales. *The Review of Economic Studies*, *78*(1), pp.356-393.
Mouganie, P. and Wang, Y. High Performing Peers and Female STEM Choices in School. *Journal of Labor Economics, 2019*
Mukherjee, P., 2015. The effects of social identity on aspirations and learning outcomes: a field experiment in rural India. *Unpublished working paper, College of William and Mary*.
Oster, E., 2019. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2), pp.187-204.
Oster, E. and Thornton, R., 2012. Determinants of technology adoption: Peer effects in menstrual cup take-up. *Journal of the European Economic Association*, *10*(6), pp.1263-1293.
Pei, Z., Pischke, J.-S., & Schwandt, H., 2018. Poorly Measured Confounders are More Useful on the Left than on the Right. *Journal of Business & Economic Statistics*, 1–12.
Pop-Eleches, C. and Urquiola, M., 2013. Going to a better school: Effects and behavioral responses. *American Economic Review*, *103*(4), pp.1289-1324.
Romano, J.P. and Wolf, M., 2005a. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469), pp.94-108.

Romano, J.P. and Wolf, M., 2005b. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), pp.1237-1282.

Sacerdote, B., 2001. Peer effects with random assignment: Results for Dartmouth roommates. *The Quarterly journal of economics*, *116*(2), pp.681-704.

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

Sacerdote, B., 2014. Experimental and quasi-experimental analysis of peer effects: two steps forward? *Annu. Rev. Econ.*, *6*(1), pp.253-272.

Salamanca, N., de Grip, A., Fouarge, D. and Montizaan, R., 2020. Locus of control and investment in risky assets. *Journal of Economic Behavior & Organization*, 177, pp.548-568.

Sapelli, C. and Illanes, G., 2016. Class size and teacher effects in higher education. *Economics of Education Review*, *52*, pp.19-28.

Sojourner, A., 2013. Identification of peer effects with missing peer data: Evidence from Project STAR. *The Economic Journal*, *123*(569), pp.574-605.

Stinebrickner, T. and Stinebrickner, R., 2001. *Peer effects among students from disadvantaged backgrounds* (No. 20013). University of Western Ontario, Centre for Human Capital and Productivity (CHCP).

Stinebrickner, R. and Stinebrickner, T.R., 2006. What can be learned about peer effects using college roommates? Evidence from new survey data and students from disadvantaged backgrounds. *Journal of public Economics*, *90*(8-9), pp.1435-1454.

Todd, P.E. and Wolpin, K.I., 2003. On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), pp.F3-F33.

Todd, P.E. and Wolpin, K.I., 2007. The production of cognitive achievement in children: Home, school, and racial test score gaps. *Journal of Human capital*, *1*(1), pp.91-136.

Todd, P. and Wolpin, K.I., 2018. Accounting for Mathematics Performance of High School Students in Mexico: Estimating a Coordination Game in the Classroom. *Journal of Political Economy*, *126*(6), pp.2608-2650.

Ushchev, P. and Zenou, Y., 2020. Social norms in networks. *Journal of Economic Theory*, *185*, p.104969.

Vardardottir, A., 2013. Peer effects and academic achievement: a regression discontinuity approach. *Economics of Education review*, *36*, pp.108-121.

Whitmore, D., 2005. Resource and peer impacts on girls' academic achievement: Evidence from a randomized experiment. *American Economic Review*, *95*(2), pp.199-203.

Wooldridge, J.M., 2007. Inverse probability weighted estimation for general missing data problems. *Journal of econometrics*, *141*(2), pp.1281-1301.

Young, A., 2019. Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2), pp.557-598.

Zimmerman, D.J., 2003. Peer effects in academic outcomes: Evidence from a natural experiment. *Review of Economics and statistics*, *85*(1), pp.9-23.

Zölitz, U. and Feld, J., forthcoming. The Effect of Peer Gender on Major Choice in Business School. *Management Science*.

APPENDICES

Appendix A. The construction of standardized scales in TEPS

We summarize the wealth of data available in TEPS into standardized summary indices using commonly used data reduction methods. We proceed as following:

1. Compute Spearman correlation of all potential variables in the factor to construct: eliminate very low correlates
2. Run preliminary PCA on remaining variables
3. Count number of missing values by individual across variables
4. Standardize each variable, construct preliminary index as row-mean across standardized variables
5. Cut preliminary index into deciles: construct bins of similar input
6. For each variable, construct median within index decile among people used for imputation (less than 1/3 missing)
7. For each variable, if missing item and less than 1/3 missing, replace missing value by median within index decile.
8. Re-run PCA now using variables with imputed values
9. Check visually that factor with and without imputed values have same distribution

In the long table below, we report for each index we use:

i) the variables used,

ii) the initial number of observations for each of these items separately,

iii) the factor loadings from the preliminary PCA, prior to imputation,

iv) the number of observations for the factor before and after imputation,

v) the eigenvalue of the first factor before and after imputation,

vi) the factor loadings from the final PCA, after imputation.

Factor for which no imputation has been performed are indicated by blanks for factor loadings after imputation, observations after imputation and eigenvalue of first factor after imputation.

Table A1. Construction of standardized scales of educational inputs in the TEPS data

		Factor loadings	
	Obs.	Original	Imputed
Effort wave 1			
Chinese teacher's assessment of student effort in class	18,508	0.75	0.76
English teacher's assessment of student effort in class	17,961	0.73	0.75
Math teacher's assessment of student effort in class	18,126	0.71	0.72
Dao Shi report student always completes homework on time	18,571	0.62	0.63
Chinese teacher's report student always completes homework on time	18,627	0.70	0.71
English teacher's report student always completes homework on time	18,233	0.67	0.68
Math teacher's report student always completes homework on time	18,394	0.65	0.66
Factor observations		16,004	19,231
First factor eigenvalue		3.35	3.46
Effort wave 2			
Chinese teacher's assessment of student effort in class	17,120	0.78	0.79
English teacher's assessment of student effort in class	16,509	0.76	0.77
Math teacher's assessment of student effort in class	16,612	0.74	0.76
Dao Shi report student always completes homework on time	17,161	0.71	0.72
Chinese teacher's report student always completes homework on time	17,107	0.68	0.69
English teacher's report student always completes homework on time	16,657	0.63	0.65
Math teacher's report student always completes homework on time	16,698	0.62	0.64
Factor observations		14,251	17,950
First factor eigenvalue		3.48	3.62
Mental health wave 1			
Self-reported frequency feeling down or frustrated	19,781	0.74	0.74
Self-reported frequency feeling troubled, worried	19,877	0.74	0.73
Self-reported frequency want to scream or smash something	19,854	0.64	0.64
Self-reported frequency feeling body shaking, unable to focus	19,839	0.68	0.68
Self-reported frequency feeling lonely	19,793	0.76	0.76
Self-reported frequency hopeless	19,856	0.75	0.75
Factor observations		19,493	19,934
First factor eigenvalue		3.09	3.09
Mental health wave 2			
Self-reported frequency feeling down or frustrated	18,716	0.71	0.71
Self-reported frequency want to scream or smash something	18,712	0.67	0.67
Self-reported frequency feeling body shaking, unable to focus	18,695	0.62	0.62
Self-reported frequency feeling lonely	18,676	0.64	0.64
Self-reported frequency feeling that you have bad fortune	18,658	0.59	0.59
Self-reported frequency feeling easily irritated by others	18,682	0.62	0.62
Self-reported frequency guilty, regret over some things	18,654	0.58	0.58
Factor observations		18,355	18,782
First factor eigenvalue		2.82	2.83
Truancy wave 1			
Self-reported frequency cutting or skipping class	19,846	0.70	0.70
Self-reported frequency physical fights or quarrels with teachers	19,790	0.61	0.6
Self-reported frequency watching porn	19,867	0.65	0.64
Self-reported frequency substance abuse (tobacco, alcohol, drugs)	19,865	0.74	0.73
Self-reported frequency running away from home	19,880	0.73	0.73
Self-reported frequency stealing or destroying others' property	19,862	0.68	0.67

	Obs.	Factor loadings	
		Original	Imputed
Factor observations		19,614	19,929
First factor eigenvalue		2.83	2.77
Truancy wave 2			
Self-reported frequency cutting or skipping class	18,718	0.51	0.52
Self-reported frequency physical fights or quarrels with teachers	18,737	0.59	0.59
Self-reported frequency watching porn	18,729	0.42	0.42
Factor observations		18,611	18,799
First factor eigenvalue		0.78	0.79
Self-efficacy wave 1			
I am good at presentations or expressing my points of view	19,749	0.65	0.65
I am good at coordinating with other people in a group	19,800	0.68	0.68
I can plan things well no matter how trivial they are	19,810	0.74	0.73
I cooperate with everyone very well	19,798	0.62	0.62
I always come up with solutions to problems	19,758	0.57	0.57
I have always reviewed what I learn since elementary school	19,847	0.59	0.59
I always try to figure out answers whenever have questions	19,808	0.59	0.59
Factor observations		19,346	19,909
First factor eigenvalue		2.83	2.83
Self-efficacy wave 2			
I am good at presentations or expressing my points of view	18,686	0.54	0.53
I am good at coordinating with other people in a group	18,744	0.58	0.58
I can plan things well no matter how trivial they are	18,731	0.65	0.64
I cooperate with everyone very well	18,709	0.55	0.54
I always come up with solutions to problems	18,708	0.62	0.62
My friends think of me as a person who always has lots of ideas	18,606	0.54	0.54
Factor observations		18,384	18,795
First factor eigenvalue		2.02	2.01
Initiative in class wave 1			
Chinese teacher's assessment of student initiative to participate in class	18,635	0.52	0.53
English teacher's assessment of student initiative to participate in class	18,307	0.52	0.54
Math teacher's assessment of student initiative to participate in class	18,367	0.52	0.54
Factor observations		17112	19219
First factor eigenvalue		0.81	0.86
Initiative in class wave 2			
Chinese teacher's assessment of student initiative to participate in class	17,161	0.58	0.61
English teacher's assessment of student initiative to participate in class	16,787	0.61	0.64
Math teacher's assessment of student initiative to participate in class	16,698	0.59	0.62
Factor observations		15,426	17,791
First factor eigenvalue		1.06	1.16
Money wave 1			
Hours per week spent on tutoring in/outside school	19,851	0.60	0.60
Amount paid for this child's tutoring classes	19,710	0.60	0.60
Factor observations		19,573	19,988
First factor eigenvalue		0.71	0.73
Money wave 2			
Hours per week spent on tutoring outside school	18,747	0.78	0.78
Monthly expenditures paid this semester for this child's tutoring classes	18,755	0.78	0.78
Factor observations		18,586	18,916

	Obs.	Factor loadings	
		Original	Imputed
First factor eigenvalue		1.21	1.22
Time wave 1			
How often parents go to bookstores or expos with child	19,750	0.53	0.53
How often parents go to concerts or performances with child	19,750	0.53	0.53
Factor observations		19,743	19,757
First factor eigenvalue		0.55	0.55
Time wave 2			
Weekly number of dinners with the child	18,783	0.44	0.45
Spouse: Weekly number of dinners with the child	18,493	0.44	0.45
Factor observations		18,457	18,819
First factor eigenvalue		0.39	0.41
Parent strictness wave 1			
My father is strict	19,851	0.51	0.51
My mother is strict	19,842	0.51	0.51
Factor observations		19,739	19,928
First factor eigenvalue		0.52	0.53
Parent strictness wave 2			
How many of your parents set strict rules for your daily routine?	18,828	0.61	0.61
How many of your parents set strict rules about spending money?	18,819	0.54	0.54
How many of your parents set strict rules about demeanor?	18,806	0.63	0.63
How many of your parents set strict rules about health habits?	18,731	0.60	0.60
How many of your parents set strict rules about making friends?	18,821	0.57	0.57
How many of your parents uses guilt and emotional blackmail?	18,821	0.51	0.51
How many of your parents does not allow you to argue with them?	18,816	0.50	0.50
How many of your parents discipline you very strictly?	18,809	0.53	0.53
Factor observations		18,648	18,831
First factor eigenvalue		2.54	2.55
Parental emotional support wave 1			
My father discusses student's future study and career	19,854	0.46	0.46
My father discusses my feelings and thoughts	19,764	0.59	0.58
My mother discusses student's future study and career	19,822	0.49	0.50
My mother discusses my feelings and thoughts	19,816	0.64	0.64
My father accepts me as I am	18,993	0.49	0.51
My mother accepts me as I am	19,370	0.49	0.49
My family provides strong emotional support	19,652	0.53	0.54
In my family, we discuss together important decisions	19,528	0.56	0.57
Factor observations		17,729	19,973
First factor eigenvalue		2.28	2.33
Parental emotional support wave 2			
My parents pay attention to my ideas and thoughts	18,816	0.66	0.66
I seek my parents' help when I encounter difficulties	18,811	0.67	0.67
My parents accept me as I am	18,799	0.62	0.62
Factor observations		18,769	18,827
First factor eigenvalue		1.27	1.27
School environment wave 1			
My school is an interesting place	19,513	0.47	0.48
My school is fair in terms of rewards and grading	19,557	0.54	0.55
The campus of my school is safe	19,567	0.63	0.63

	Obs.	Factor loadings	
		Original	Imputed
My school cares about their students	19,481	0.71	0.71
My school has a great atmosphere for learning	19,456	0.64	0.65
Factor observations		18,701	19,903
First factor eigenvalue		1.83	1.86
School environment wave 2			
My school's requirements on students are quite reasonable	18,614	0.39	0.39
My school is fair in terms of rewards and grading	18,741	0.46	0.46
The campus of my school is safe	18,709	0.56	0.56
My school cares about their students	18,340	0.62	0.62
My school has a great atmosphere for learning	18,690	0.52	0.52
Factor observations		18,053	18,814
First factor eigenvalue		1.33	1.34
Teacher engagement wave 1			
How many of my teachers know the name of every student	19,865	0.38	0.39
How many teachers encourage student when they study hard	19,780	0.48	0.48
How many teachers use different teaching methods/materials	19,846	0.55	0.55
How many teachers give homework to increase students' chance to practice	19,836	0.48	0.49
How many teachers ask reasons when students fail on homework	19,812	0.46	0.48
How many teachers give a review after every exam	19,604	0.48	0.49
Factor observations		19,210	19,953
First factor eigenvalue		1.35	1.4
Teacher engagement wave 2			
How many teachers talk about people skills in class	18,795	0.70	0.70
How many teachers often discuss life goals/conduct career planning in class	18,784	0.73	0.73
How many teachers often recommend good books and encourage reading in class	18,783	0.62	0.62
How many teachers often use real life and practical examples in class	18,772	0.62	0.62
How many teachers take on his spare time to talk to students who have personal problems	18,795	0.53	0.53
How many teachers often use guilt or emotional blackmail	18,784	0.45	0.45
How many teachers praise me when I study hard	18,744	0.53	0.53
Factor observations		18,590	18,820
First factor eigenvalue		2.56	2.56

Appendix B. Additional Tables and Figures

Table B1. The effect of better peer test scores on students' own test scores in wave 2

Outcome:	**Student test scores in wave 2 [std]**			
Peer test scores [std]	0.042**	0.047***	0.054***	0.054***
	(0.017)	(0.017)	(0.017)	(0.017)
Own test scores [std]	0.708***	0.703***	0.566***	0.562***
	(0.007)	(0.007)	(0.008)	(0.009)
R^2	0.64	0.64	0.68	0.68
School FE	✓	✓	✓	✓
Balancing controls		✓		✓
W1 inputs			✓	✓
Schools	227	227	227	227
Classrooms	853	853	853	853
Students	12816	11925	11734	11068

This table reports estimates of regressing standardized student test scores in wave 2 on standardized average peer test scores in wave 1 in our sample containing 227 schools, 853 classrooms, and up to 12,816 students. Balancing controls include household income, family engagement with homework, gifted art classroom assignment, and parents' efforts to get child assigned to a particular classroom. W1 inputs include standardized scales of student inputs (school effort, initiative in class, truancy, academic self-efficacy, and mental health), parent inputs (investment in private tutoring, time investments, parental strictness and parental support), school and teacher inputs (school environment and teacher engagement). Standard errors clustered at the classroom level in parentheses. Estimates in this figure are also shown in Figure 3.

Table B2. The effect of better peer test scores on educational inputs in wave 2

Treatment:	Peer test scores [std]				
	Coef.	Std. err.	R^2	Classrooms	Students
Outcomes: educational inputs					
School effort [std]	-0.058**	(0.026)	0.56	852	10,694
Initiative in class [std]	-0.022	(0.026)	0.45	852	10,588
Truancy [std]	-0.002	(0.024)	0.18	853	11,151
Cheated on exams [.48]	0.010	(0.014)	0.12	853	11,116
Academic self-efficacy [std]	-0.020	(0.023)	0.15	853	11,151
Mental health [std]	-0.028	(0.022)	0.16	853	11,141
University aspirations [.57]	0.017*	(0.010)	0.28	853	11,153
University expectations [.44]	0.021**	(0.010)	0.29	853	11,143
Private tutoring [std]	-0.000	(0.019)	0.37	853	11,204
Time with parents [std]	0.077***	(0.024)	0.08	853	11,148
Conflict with parents [.31]	-0.016	(0.011)	0.06	853	11,121
Parental strictness [std]	0.030	(0.022)	0.16	853	11,171
Parental support [std]	0.033	(0.022)	0.20	853	11,171
Harsh parenting [.33]	0.015	(0.010)	0.08	853	11,171
Parent uni. aspirations [.51]	0.006	(0.011)	0.33	853	11,058
School environment [std]	-0.032	(0.025)	0.17	853	11,164
Teacher engagement [std]	0.017	(0.026)	0.11	853	11,167

This table reports estimates of regressing educational input measures in wave 2 on standardized average peer test scores in wave 1 in our sample containing 227 schools, 853 classrooms, and up to 11,067 students. Rows present results of models with different educational inputs as outcomes. Unconditional means of each outcome are shown in square brackets, and [std] marks outcomes that have been standardized to have a mean of zero and a standard deviation of one. All models control for school fixed effects, student test scores in wave 1, balancing controls, and educational inputs in wave 1. Standard errors clustered at the classroom level in parentheses. Estimates in this table are also shown in Figure 4.

Table B3. Returns to educational inputs from cumulative value-added models

Outcome:	**Student test scores in wave 2 [std]**			
	Total effect		Partial effect	
	Coef.	Std. err.	Coef.	Std. err.
Treatment:				
School effort [std]	0.158***	(0.009)	0.093***	(0.009)
Initiative in class [std]	0.161***	(0.008)	0.109***	(0.008)
Truancy [std]	-0.036***	(0.006)	-0.008	(0.006)
Cheated on exams [.48]	-0.053***	(0.011)	-0.015	(0.011)
Academic self-efficacy [std]	-0.007	(0.006)	-0.023***	(0.006)
Mental health [std]	0.008	(0.006)	-0.001	(0.006)
University aspirations [.57]	0.147***	(0.013)	0.047***	(0.013)
University expectations [.44]	0.204***	(0.013)	0.125***	(0.014)
Private tutoring [std]	0.041***	(0.007)	0.027***	(0.007)
Time with parents [std]	-0.003	(0.006)	-0.003	(0.006)
Conflict with parents [.31]	0.070***	(0.012)	0.038***	(0.012)
Parental strictness [std]	-0.029***	(0.006)	-0.039***	(0.006)
Parental support [std]	0.029***	(0.006)	0.019***	(0.006)
Harsh parenting [.33]	-0.056***	(0.011)	-0.025**	(0.012)
Parent uni. aspirations [.51]	0.135***	(0.013)	0.068***	(0.012)
School environment [std]	0.025***	(0.006)	0.007	(0.006)
Teacher engagement [std]	0.014**	(0.006)	0.005	(0.006)
R^2			0.71	
Schools			227	
Classrooms			852	
Students			10,771	

This table reports coefficient estimates of regressing student test scores in wave 2 on educational inputs in wave 2 in our estimation sample containing 227 schools, up to 853 classrooms, and up to 12,816 students. Rows present coefficients of different regressors. Unconditional means of each input are shown in square brackets and [std] marks inputs that have been standardized to have a mean of zero and a standard deviation of one. Total effects are estimated one input at the time, whereas partial effects are estimates of all inputs jointly. All models control for school fixed effects, student test scores, average peer test scores, and educational inputs in wave 1. Standard errors clustered at the classroom level in parentheses. These results are also available in Figure 5.

Table B4. Academic peer effect mediated by educational inputs

Outcome:	Student test scores in wave 2 [std]	
	Coef.	Std. err.
Total mediated effect	-0.009*	(0.005)
Mediated effect by		
School effort	-0.006***	(0.002)
Initiative in class	-0.003	(0.003)
Truancy	-0.000	(0.000)
Cheated on exams	-0.000	(0.000)
Academic self-efficacy	0.001	(0.001)
Mental health	0.000	(0.000)
University aspirations	0.000	(0.000)
University expectations	0.001	(0.001)
Private tutoring	-0.000	(0.001)
Time with parents	-0.000	(0.000)
Conflict with parents	-0.000	(0.000)
Parental strictness	-0.000	(0.001)
Parental support	0.000	(0.000)
Harsh parenting	-0.000	(0.000)
Parent uni. Aspirations	-0.000	(0.001)
School environment	-0.000	(0.000)
Teacher engagement	0.000	(0.000)

This table reports the mediated effects based on Gelbach's (2016) decomposition of our academic peer effect estimate using only within-school variation in our estimation sample containing 227 schools, up to 853 classrooms, and up to 12,816 students. These estimates are produced using a modified version of the `b1x2` *Stata package. Rows present the mediated effect of different educational inputs in wave 2. All models control for school fixed effects, student test scores, average peer test scores, and educational inputs in wave 1. Standard errors clustered at the classroom level in parentheses. These results are also available in Figure 6.*

Table B5. Permutation-based class-level sorting tests in estimation sample

		Share of classrooms with empirical p-values under			
	Classrooms	0.10	0.05	0.01	Avg. p-value
Pre-assignment characteristics					
Student test scores	853	0.10	0.06	0.02	0.486
Female student	853	0.06	0.04	0.02	0.562
Student born before 1989	853	0.10	0.05	0.01	0.490
Monthly household income over NT$100,000	853	0.09	0.04	0.01	0.491
College-educated parent(s)	853	0.09	0.06	0.02	0.485
Parent(s) work in government	853	0.08	0.04	0.01	0.487
Ethnic minority parent(s)	853	0.08	0.04	0.01	0.494
Student prioritized studies since primary school	853	0.12	0.06	0.01	0.491
Student reviews lessons since primary school	853	0.12	0.06	0.01	0.478
Student likes new things since primary school	853	0.13	0.07	0.02	0.465
Student was truant in primary school	853	0.08	0.04	0.01	0.498
Student had mental health issues in primary school	853	0.10	0.06	0.02	0.495
Had private tutoring before junior high school	853	0.11	0.06	0.01	0.479
Family help with homework before junior high school	853	0.09	0.05	0.01	0.496
Student quarreled with parents in primary school	853	0.10	0.05	0.01	0.503
Student enrolled in gifted academic class	853	0.09	0.05	0.02	0.466
Student enrolled in arts gifted class	853	0.12	0.08	0.03	0.447
Parents made efforts to place student in better class	853	0.13	0.07	0.02	0.465

This table shows the results of permutation-based class-level sorting tests, in our estimation sample containing 227 schools, 853 classrooms, and 12,816 students. For these tests, we simulate 10,000 classrooms under the null of random assignment of students to classrooms within schools, calculate the mean of pre-treatment characteristics in synthetic classroom, and construct class-level empirical p-values as the share of times synthetic classroom means were more extreme than actual classroom means relative to the schools mean. Each row presents class-level empirical p-values for a different pre-assignment characteristic. The last column shows the average p-value for all classrooms.

Table B6. Non-parametric sorting test in estimation sample

	No. of school-level regressions	Share of class-dummy joint significance test p-values under .10	.05	.01
Outcomes: Pre-assignment characteristics				
Student test scores	227	0.06	0.04	0.03
Female student	216	0.05	0.02	0.02
Student born before 1989	227	0.12	0.03	0.01
Monthly household income over NT$100,000	208	0.09	0.04	0.00
College-educated parent(s)	204	0.13	0.07	0.02
Parent(s) work in government	205	0.06	0.02	0.01
Ethnic minority parent(s)	179	0.06	0.02	0.01
Student prioritized studies since primary school	227	0.12	0.06	0.01
Student reviews lessons since primary school	227	0.10	0.06	0.02
Student likes new things since primary school	227	0.14	0.10	0.02
Student was truant in primary school	227	0.10	0.03	0.01
Student had mental health issues in primary school	227	0.12	0.07	0.01
Had private tutoring before junior high school	227	0.13	0.08	0.02
Family help with homework before junior high school	226	0.08	0.06	0.02
Student quarreled with parents in primary school	227	0.10	0.04	0.00
Student enrolled in gifted academic class	206	0.11	0.05	0.02
Student enrolled in arts gifted class	186	0.15	0.09	0.07
Parents made efforts to place student in better class	225	0.14	0.10	0.04

This table shows the results of non-parametric school-level sorting tests in our estimation sample containing 227 schools, 853 classrooms, and 12,816 students. School-by-school, we regress each pre-treatment characteristics on a set of class dummies, F-test them for joint significance, and calculate the share of times the F-tests p-values fall under typical significance thresholds. Each row presents class-level empirical p-values for a different pre-assignment characteristic. We use cluster-robust covariance matrices at the classroom level for each test.

Table B7. Oster (2019) proportional selection on unobservables in initial sample

	Degree of selection required to explain effect of peer test scores on outcomes
Outcomes:	
Test scores	-0.20
School effort	-0.10
Initiative in class	-1.30
Truancy	-0.30
Cheated on exams	-0.70
Academic self-efficacy	-1.80
Mental health	-0.60
University aspirations	-0.90
University expectations	-1.20
Private tutoring	0.10
Time with parents	-0.40
Conflict with parents	-3.60
Parental strictness	-1.70
Parental support	-0.50
Harsh parenting	-4.40
Parent uni. Aspirations	-0.20
School environment	-0.70
Teacher engagement	-0.90
Selection proportional to:	
Balancing controls	✓
W1 inputs	✓

This table reports Oster's (2019) δ, the share of proportional selection needed to explain away each estimate in our initial sample 332 schools, 1,241 classrooms and 14,383 students. Values of δ between zero and one imply that, under reasonable assumption, the effect can be explained by correlated unobservables. Each cell is an estimate from a separate analysis. All estimates are calculated using Oster's (2019) `psacalc` *Stata package, and assume a theoretical maximum R-square of one. All models control for school fixed effects and student test scores in wave 1. Pre-assignment characteristics are listed in Section 3.4. Educational inputs in wave 1 are listed in Section 4.1*

Table B8. The effect of better peer ability on students' own ability using alternative measures of ability

Outcome: Student ability in wave 2 [std]					
	Ability measure used:				
			IRT Bayesian posterior mean of		
	Analytical	Mathematical	General	Analytical	Mathematical
Peer ability [std]	0.039**	0.048***	0.052***	0.040**	0.051***
	(0.019)	(0.017)	(0.018)	(0.020)	(0.018)
Own ability [std]	0.385***	0.535***	0.592***	0.394***	0.555***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
R^2	0.46	0.61	0.70	0.49	0.64

*This table reports coefficient estimates of regressing student's own ability in wave 2 on standardized average peer ability and own ability in wave 1 in our estimation sample containing 227 schools, 853 classrooms, and 12,816 students. The columns vary the measure of ability used for the analysis. The identification of analytical and mathematical subcomponents of ability and the Bayesian posterior mean calculation based on Item Response Theory (IRT) models, the TEPS team could also identify two highly correlated but distinct subcomponents measuring analytical ability and mathematical ability based on disjoint subsets of test questions. The IRT models were also used to produce the standardized Bayesian posterior means of the three components identifiable in the test—the general ability component and the analytical ability and mathematical ability subcomponents. All models include school fixed effects and educational inputs in wave 1. Standard errors are clustered at the classroom level. *, ** and *** denote significance levels at the 10%, 5% and 1%.*

Table B9. The effect of better peer ability on students' own ability using instrumental variable estimators to account for measurement error in ability

Outcome: Student ability in wave 2 [std]			
	Measure of ability		
	Analytical	Mathematical	Mixed
Peer ability [std]	0.050*	0.038	0.039
	(0.030)	(0.027)	(0.030)
Instrument	Mathematical	Analytical	Alt. mixed
t-statistic of first-stage coefficient	30.64	28.24	26.79

*This table reports coefficient estimates of instrumental variable regressions of student's test scores in wave 2 on standardized average peer ability in wave 1 in our estimation sample containing 227 schools, 853 classrooms, and 12,816 students. The measures of ability and the instrument vary across columns, as described in Section 5.2.2. All models include school fixed effects, and students' own test scores and educational inputs in wave 1. Standard errors are clustered at the classroom level. *, ** and *** denote significance levels at the 10%, 5% and 1%.*

Table B10. The effect of peer ability on educational inputs using a mixed ability IV approach

	Mixed IV effect of peer ability [std]	
	Coef.	Std. err.
Outcomes: educational inputs		
School effort [std]	-0.083**	(0.035)
Initiative in class [std]	-0.035	(0.037)
Truancy [std]	0.018	(0.035)
Cheated on exams [.48]	0.015	(0.020)
Academic self-efficacy [std]	0.001	(0.033)
Mental health [std]	-0.048	(0.034)
University aspirations [.57]	0.029**	(0.014)
University expectations [.44]	0.029**	(0.015)
Private tutoring [std]	0.018	(0.027)
Time with parents [std]	0.073**	(0.036)
Conflict with parents [.31]	-0.013	(0.017)
Parental strictness [std]	0.053	(0.034)
Parental support [std]	0.032	(0.033)
Harsh parenting [.33]	0.006	(0.015)
Parent uni. aspirations [.51]	0.004	(0.017)
School environment [std]	-0.033	(0.038)
Teacher engagement [std]	0.005	(0.037)

*This table reports coefficient estimates of instrumental variable regressions of student's educational inputs in wave 2 on standardized average peer ability in wave 1 in our estimation sample containing 227 schools, 853 classrooms, and 12,816 students. Peer ability and its instrument are constructed using the 'mixed IC' approach described in Section 5.2.2. All models include school fixed effects, and students' own ability and educational inputs in wave 1. Standard errors are clustered at the classroom level. *, ** and *** denote significance levels at the 10%, 5% and 1%.*

Table B11. The effect of peer ability using Sojourner's (2013) correction for incomplete class sampling

	Effect of peer test scores [std] with Sojourner (2013) correction for peer test scores missing not at random					
Outcomes:						
Test scores [std]	0.126***	0.090**	0.089**	0.090**	0.092***	0.097***
	(0.039)	(0.037)	(0.037)	(0.036)	(0.035)	(0.036)
School effort [std]	-0.041	-0.039	-0.038	-0.045	-0.051	-0.054
	(0.059)	(0.055)	(0.055)	(0.055)	(0.054)	(0.054)
Initiative in class [std]	-0.085	-0.065	-0.060	-0.072	-0.081	-0.082
	(0.066)	(0.058)	(0.057)	(0.057)	(0.056)	(0.056)
Truancy [std]	-0.048	-0.020	-0.020	-0.018	-0.016	-0.021
	(0.060)	(0.058)	(0.058)	(0.059)	(0.059)	(0.058)
Cheated on exams [.48]	0.035	0.015	0.020	0.030	0.029	0.026
	(0.035)	(0.031)	(0.030)	(0.031)	(0.031)	(0.030)
Academic self-efficacy [std]	-0.023	-0.029	-0.021	-0.012	-0.003	0.003
	(0.056)	(0.050)	(0.049)	(0.050)	(0.049)	(0.049)
Mental health [std]	-0.015	-0.009	-0.009	-0.008	0.002	0.002
	(0.056)	(0.049)	(0.049)	(0.050)	(0.049)	(0.049)
University aspirations [.57]	0.053**	0.040*	0.042*	0.040*	0.041*	0.038*
	(0.025)	(0.022)	(0.022)	(0.022)	(0.021)	(0.021)
University expectations [.44]	0.050**	0.031	0.032	0.029	0.034	0.033
	(0.025)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
Private tutoring [std]	-0.003	0.004	-0.006	0.010	0.028	0.025
	(0.048)	(0.043)	(0.042)	(0.043)	(0.042)	(0.041)
Time with parents [std]	0.064	0.137***	0.135***	0.118**	0.152***	0.147***
	(0.056)	(0.049)	(0.049)	(0.048)	(0.049)	(0.049)
Conflict with parents [.31]	0.001	-0.008	-0.006	-0.008	-0.014	-0.015
	(0.027)	(0.023)	(0.023)	(0.023)	(0.022)	(0.022)
Parental strictness [std]	-0.000	0.020	0.033	0.027	0.033	0.032
	(0.058)	(0.050)	(0.050)	(0.049)	(0.048)	(0.048)
Parental support [std]	0.042	0.051	0.064	0.060	0.067	0.064
	(0.051)	(0.046)	(0.046)	(0.046)	(0.045)	(0.045)
Harsh parenting [.33]	0.019	0.006	0.008	0.015	0.009	0.006
	(0.024)	(0.022)	(0.022)	(0.022)	(0.021)	(0.021)
Parent uni. aspirations [.51]	0.037	0.017	0.016	0.012	0.014	0.015
	(0.029)	(0.024)	(0.024)	(0.024)	(0.023)	(0.023)
School environment [std]	0.012	-0.050	-0.050	-0.046	-0.065	-0.057
	(0.069)	(0.059)	(0.058)	(0.058)	(0.057)	(0.056)
Teacher engagement [std]	0.037	-0.002	0.005	-0.016	0.008	0.011
	(0.065)	(0.057)	(0.058)	(0.057)	(0.056)	(0.057)
Share of peers observed × School FE	✓					
Share of peers observed × School K-cile FE		25	20	15	10	5

This table reports coefficient estimates of regressing student outcomes in wave 2 on standardized average peer ability in wave 1 in our estimation sample containing 227 schools, 853 classrooms, and 12,816 students. These estimates correct for peer test scores missing not at random following Sojourner (2013) and implemented using Correia's (2018) `reghdfe` *Stata package. All models include school fixed effects, and students' own ability and educational inputs in wave 1. Standard errors are clustered at the classroom level. *, ** and *** denote significance levels at the 10%, 5% and 1%.*

Table B12. Corrected p-values for the effect of peer ability using Young's (2019) randomization inference and Romano and Wolf's (2007) step-down familywise error rate adjustment procedures

	Corrected p-values for the effect of peer test scores [std] using	
	Young's (2019) Randomization-t inference	Romano and Wolf's (2005) step-down procedure
Outcomes:		
Test scores	**0.002**	**0.050**
School effort	*0.052*	0.424
Initiative in class	0.433	0.958
Truancy	0.950	1.000
Cheated on exams	0.537	0.958
Academic self-efficacy	0.447	0.958
Mental health	0.255	0.868
University aspirations	0.170	0.760
University expectations	*0.064*	0.468
Private tutoring	0.994	1.000
Time with parents	**0.006**	*0.054*
Conflict with parents	0.196	0.844
Parental strictness	0.202	0.858
Parental support	0.146	0.828
Harsh parenting	0.196	0.844
Parent uni. Aspirations	0.627	0.958
School environment	0.251	0.868
Teacher engagement	0.537	0.958

This table corrected p-values for our main results using i) Young's (2019) randomization-t inference procedure to account for high-leverage, finite sample properties of the model error term, and the complex sampling structure of our data (Col. (1) based on 499 permutations), and ii) Romano and Wolf's (2007) step-down procedure for controlling for familywise error rate in multiple hypotheses testing implemented using Clarke et al.'s (2019) `rwolf` *Stata package (Col. (2), based on 499 replications). p-values smaller than 0.10 are shown in italics and smaller than 0.05 in bold.*

Table B13. Heterogeneous and mediated effects of peer ability

	Academic peer effect	Mediated effect			
		Total	by student inputs	by parent inputs	by school inputs
by monthly household income:					
Less than NT$20,000	0.055*	-0.030**	-0.016	-0.011	-0.002
	(0.028)	(0.014)	(0.012)	(0.008)	(0.003)
NT$20,000 to NT$50,000	0.057***	0.001	-0.000	0.001	-0.000
	(0.019)	(0.007)	(0.006)	(0.002)	(0.000)
NT$50,000 to NT$100,000	0.052***	-0.010	-0.009	-0.001	-0.000
	(0.020)	(0.008)	(0.008)	(0.003)	(0.000)
More than NT$100,000	0.052**	-0.019	-0.016	-0.005	0.002
	(0.025)	(0.014)	(0.013)	(0.007)	(0.002)
by parent(s) education:					
No college degree	0.050***	-0.008	-0.007	-0.001	0.000
	(0.017)	(0.005)	(0.005)	(0.002)	(0.000)
College degree	0.043*	-0.015	-0.011	-0.005	0.000
	(0.025)	(0.016)	(0.014)	(0.006)	(0.001)
by student test scores:					
Bottom tertile	0.034*	-0.005	-0.003	-0.002	-0.000
	(0.018)	(0.007)	(0.006)	(0.003)	(0.001)
Middle tertile	0.073***	-0.010	-0.006	-0.005	0.001
	(0.019)	(0.009)	(0.007)	(0.003)	(0.001)
Top tertile	0.049**	-0.017*	-0.014*	-0.002	-0.001
	(0.020)	(0.009)	(0.008)	(0.003)	(0.001)
by student gender:					
Male	0.056***	-0.014*	-0.009	-0.004*	0.000
	(0.019)	(0.007)	(0.006)	(0.002)	(0.000)
Female	0.057***	-0.006	-0.008	0.003	-0.000
	(0.018)	(0.008)	(0.007)	(0.002)	(0.001)
by school type:					
Public		-0.009*	-0.007	-0.002	-0.000
		(0.005)	(0.005)	(0.002)	(0.000)
Private		-0.004	-0.004	-0.003	0.004
		(0.014)	(0.012)	(0.004)	(0.004)
by Dao Shi experience:					
10 years or less	0.060***	-0.019**	-0.020**	-0.001	0.002
	(0.022)	(0.009)	(0.008)	(0.002)	(0.001)
More than 10 years	0.044**	-0.005	-0.003	-0.002	0.000
	(0.018)	(0.006)	(0.006)	(0.002)	(0.000)

This table reports peer and mediated effects based on Gelbach's (2016) decomposition using only within-school variation in our estimation sample containing 227 schools, 853 classrooms, and 12,816 students. These estimates are produced using a modified version of the `b1x2` *Stata package. Rows present the peer and mediated effects for different subgroups defined based on wave 1 variables. All models control for school fixed effects, student test scores, average peer test scores, and educational inputs in wave 1. Standard errors are clustered at the classroom level. *, ** and *** denote significance levels at the 10%, 5% and 1%.*

Appendix C. The fishing algorithm

In this appendix we explain the steps of our fishing algorithm introduced in Section 3.2 in detail. We illustrate its use in the TEPS data. In Appendix D we provide Monte Carlo style evidence of its performance in simulated data.

Sorting of students into classrooms within schools in TEPS

Taiwan has an explicit mandate of random assignment of students to classrooms within schools. We first test whether the TEPS data is consistent with this mandate without imposing any sample restrictions and refer to this as our "initial sample". This initial sample includes a total of 20,055 students assigned to 1,244 classrooms across 333 schools in wave 1, for whom we have data from either students, parents, teachers or school administrators' questionnaires. Most students can be matched across questionnaires—we lose fewer than 1,000 observations due to questionnaire non-match—yet we estimate our initial tests on this unrestricted sample to limit the influence of selective questionnaire attrition.

We first run sorting tests on student wave 1 standardized test scores, as well as on each characteristics we can unambiguously treat as pre-assignment; that is, variables capturing either fixed traits or events prior to entering junior high school.

Standardized test scores are not strictly measured pre-assignment; they were taken by students during the first weeks of the first junior high school academic year, shortly after assignment to classrooms. However, it is highly doubtful that only a few weeks' worth of exposure to peers could generate considerable peer effects already. Moreover, these test scores were never revealed to students, parents, teachers or school administrators so there is no chance of re-sorting of classrooms after initial assignment based on the results of these exams. However, finding sorting on standardized test scores would still be consistent with students being assigned to classrooms based on other ability or academic performance measures that are either known to the parents, teachers, or school administrators. In this spirit, we analyze standardized test scores in this paper.

To run sorting tests loosely follow the within-school equation

$$Y_{ics1} = \beta \bar{Y}_{ics1}^{-i} + \mu_s + \varepsilon_{ics1}, \qquad \text{(C1)}$$

where Y_{ics} is the characteristic of student i in class c in school s in wave 1, which is pre-determined at the time of assignment, $\overline{Y}^{-i}_{ics1}$ is the class leave-out mean of the same variable Y at wave 1 (the classroom peer mean of characteristic Y), μ_s is school-invariant unobserved heterogeneity which we account for using school fixed effects, and ε_{ics1} is a conditionally uncorrelated model error term.

The sorting statistic of interest is closely related to $\hat{t} = \hat{\beta}/\text{std. err.}(\hat{\beta})$ with the critical values of the standard normal distribution as reference in large samples. A positive $\hat{t}$ over critical values in the distribution indicates positive sorting of students into classrooms based on the tested pre-determined characteristic. However, Guryan, Kroft and Notowidgo (2009) observe that, under random assignment, $\hat{\beta}$ present a small negative bias which seems to disappear when controlling for school-level leave-out-mean of the characteristic in sorting tests. Jochmans (2020) argues that Guryan, Kroft and Notowidgo's empirical correction results in low power for detecting sorting, derives analytical expressions for this bias in within-school estimators and proposes a bias-corrected $\hat{t}$ *that* solves this power issue. In our sorting tests, we present $\hat{t}$ using the more commonly found Guryan, Kroft and Notowidgo (2009) method and the very recent Jochmans (2020) improvement.

The second and third columns of Table C1 show the sorting test statistics for all pre-determined characteristics we consider. There is plenty of evidence suggesting that students are sorted into classrooms with similar peers in the initial sample: certainly for test scores, but also for family income and parental education, intellectual curiosity during primary school, private tutoring before entering junior high school, gifted academic and art class assignment, and on parents' efforts to influence the student's classroom assignment. Sorting on test scores in this sample is already reason enough for thinking that estimates of higher-ability peer effects might be biased. Yet further balancing tests on higher-ability peers—which regress Y_{ics1} on $\overline{\text{Test Scores}}^{-i}_{ics1}$—also show that

Table C1. Balancing and Sorting Tests on Initial Sample

Treatment:		Sorting tests (t-statistic)		Balancing tests	
		Peer outcome leave-out-mean		Peer ability leave-out-mean [std]	
	Students	Guryan et al. (2009)	Jochmans (2020)	Coef.	Std. err.
Outcomes: Pre-assignment characteristics					
Student test scores [std]	19,957	**3.0**	**6.6**		
Female student	19,957	*2.2*	-0.9	0.012	(0.007)
Student born before 1989	19,866	-0.4	1.4	-0.011**	(0.005)
Household income > NT$100k/mo.	19,629	0.9	*2.2*	0.014***	(0.004)
College-educated parent(s)	19,073	1.1	**3.5**	0.036***	(0.005)
Parent(s) work in government	18,979	1.3	*2.2*	0.024***	(0.004)
Ethnic minority parent(s)	19,070	1.5	1.9	-0.011***	(0.004)
Prioritized studies since primary school	19,830	*-2.1*	1.5	-0.006	(0.005)
Reviews lessons since primary school	19,813	0.0	**2.6**	-0.002	(0.004)
Likes new things since primary school	19,771	1.0	**2.9**	0.005	(0.006)
Was truant in primary school	19,674	1.3	0.4	-0.022***	(0.005)
Student had mental health issues in primary school	19,670	0.0	0.3	0.001	(0.006)
Had private tutoring before junior high	19,720	1.5	**2.5**	0.024***	(0.006)
Family help with homework before junior high	18,976	1.3	1.2	0.006	(0.004)
Student quarreled with parents in primary school	19,691	-0.5	-1.0	-0.006	(0.006)
Student enrolled in gifted academic class	19,779	**2.3**	**4.3**	0.074***	(0.009)
Student enrolled in arts gifted class	19,779	**4.8**	**5.5**	0.033***	(0.010)
Parents made efforts to place student in better class	19,698	**5.8**	**4.8**	0.050***	(0.006)

*Estimates in our trimmed sample of 333 schools and 1,257 classrooms. All estimators include school fixed effects. The reference distribution for the Guryan et al. (2009) and the Jochmans (2020) sorting statistics is the standard normal. t-statistics larger than critical values for a two-sided test are shown in italics for 95% confidence and in bold for 99% confidence. The last column reports cluster-robust standard errors at the classroom level. ***, ** and * mark estimates statistically different from zero at the 90, 95 and 99 percent confidence level.*

higher-ability peers are also related to several pre-determined characteristics at baseline. These balancing test results are shown in the last two columns of Table C1.[8]

Our next step is to characterize the deviations from random assignment in this initial sample in order to hopefully correct them. In Taiwan, class assignment is tasked to schools themselves, as opposed to being done at the regional or school district level. Because of this, we suspect that

[8] Note that, due to the large number of pre-treatment characteristics we test and the many students and classes in TEPS, we are more likely to find imbalances than many previous academic peer effect studies. The size of our detected imbalances is relatively small generally (very) small. In fact, simple back-of-the-envelope calculations suggest that in other datasets commonly used to estimate peer effects, such as the Project STAR data, imbalances of this size would have gone undetected.

deviations from random assignment in our data could come directly from having non-compliant schools, and direct our efforts towards finding these schools. All results in Table C1 suggest that, in these defier schools, students assigned to higher-ability peers are also higher ability themselves and are also generally more advantaged in other respects. These schools might have sorted students into classrooms directly based on academic ability/performance, perhaps by assignment them to "gifted" classrooms together, and perhaps also as a response of parental pressure on the school. All these are informative insights in the next steps of our fishing algorithm.

The Fishing Algorithm

The Fishing Algorithm is a data-driven method we developed to detect schools that are likely not compliant with Taiwan's national mandate to randomly assign students to classrooms. The algorithm combines permutation-based measures of the degree of sorting in the data with latent-class modeling techniques. Despite seeming complex, the intuition behind the procedure is simple and its implementation is fast. Its steps are described in Box C1.

Box C1. The Fishing Algorithm

Step 1	Identify sorted/imbalanced pre-assignment characteristics. Identify your key measure of interest and, if sorted/imbalanced, continue to step 2.
Step 2	Construct a school-level measure of sorting in your key measure of interest for each school $s = 1 \ldots S$. We propose a modified Herfindahl-Hirschman index for concentration of the key student characteristic into classrooms in each school. Call this measure H_s.
Step 3	For each school, simulate the counterfactual H_s under random assignment of students to classrooms, while keeping school size, class size, number of classrooms and student compositions constant. Call this counterfactual assignment H_s^{random}. Use B permuted random assignments of students to classrooms to derive the school-level distributions of H_s^{random} for each school s. Using these distributions, construct the school-level share of permutations for which H_s is larger than H_s^{random} and call it $S_s \in [0,1]$. S_s measures of the degree of sorting of students to classrooms in each school over and above what chance would predict.
Step 4	Use latent class models to predict S_s. Since S_s is censored below at 0 and above at 1, we propose fitting finite mixture tobit regressions. Select the number of latent classes in the model using a pre-determined goodness-of-fit measure (e.g., AIC, BIC). (If available, use school-level predictors for defier schools informed by your knowledge of the data. You can use likelihood ratio tests to decide whether these class predictors are worth including in the model.) Identify the latent class(es) associated with high S_s (close to 1); these are likely to capture defier schools. Using model estimates, construct the school-level posterior probability of belonging to a defier class. Call this measure P_s.
Step 5	Construct a "likely defier" flag for each school based on whether P_s exceeds a pre-determined threshold. We suggest using a "most likely defier" rule: flag schools which are most likely to belong to a defier class as defier schools. Remove flagged schools from your estimation sample, call this the trimmed sample. Re-estimate your balancing/sorting tests in this sample.

In the first step, we identify whether there is evidence of sorting and/or imbalance in the data. Table C1 describes the results of these tests for the TEPS initial sample. Since our study focuses on estimating the effect of higher-ability classroom peers, we identify student test scores as our key pre-assignment characteristic for the remaining steps.

In the second step, we construct our school-level measure of sorting of students into classrooms based on standardized test scores. We base our measure on the Herfindahl-Hirschman index, the most prominent measure of market concentration in economics. In school s with classrooms $c = 1 \dots C$, we define our measure as

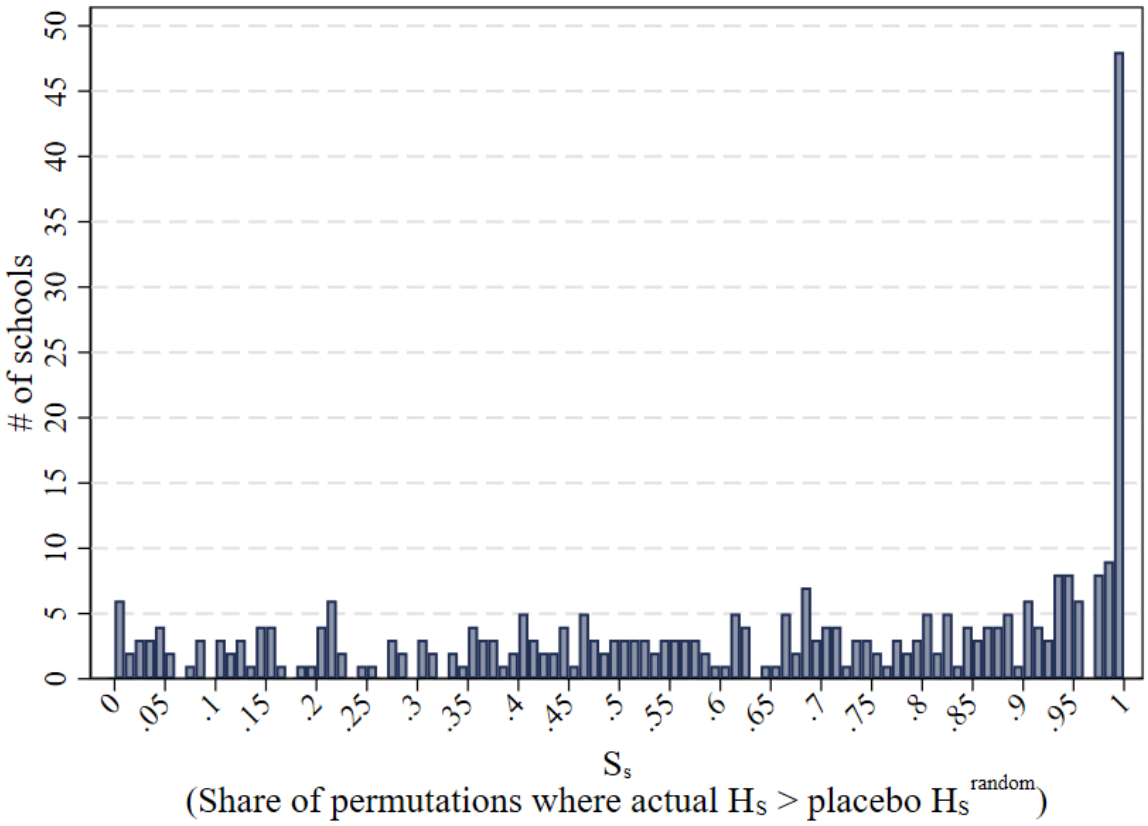
$$H_s = \sum_{c=1}^{C} \left(\frac{\overline{\text{Test scores}}_{cs}}{\sum_{c=1}^{C} \overline{\text{Test scores}}_{cs}} \right)^2, \tag{C2}$$

where $\overline{\text{Test scores}}_{cs}$ is the average standardized test score in classroom c of school s.[9] H_s is a measure of the concentration (or sorting) of student test scores into classrooms within each school, and it will range between $1/C$ (if $\overline{\text{Test scores}}_{cs}$ is identical in all classrooms) to 1 (if all students with positive test scores are together in one classroom—which is ridiculous with test scores but more easy to think of when measuring sorting by e.g., race or gender). At this point our constructed H_s includes test score sorting data for each of the 333 schools in TEPS.

In the third step, we construct counterfactual distributions of H_s for each of the 333 schools in TEPS that reflect random assignment of students to classrooms within schools. To do this, we construct $B = 400$ permutations of random assignment of students to classroom within each school maintaining each schools' data structure; that is, maintaining the student number and composition in each school, and the exact number and size of classrooms in each school. Ensuring the data structure is maintained is crucial for computing randomization-based statistics (Young, 2019). For each permutation $b = 1 \dots B$ we thus end up with a measure H_s^{random} that reflects one way school sorting *could have looked like* if classrooms were randomly assigned within schools.

[9] It is important to note that by standardized test scores we mean "scores from a standardized test" rather than "test scores that have been standardized to have a mean of zero and a standard deviation of one". Steps 3 through 5 of the fishing algorithm work much better if H_s is constructed from test scores (or any other measure) that is weakly positive (i.e., with support in $[0, \infty)$).

Figure C1. The school-level concentration of classrooms based on test scores, S_s, in the TEPS data based on $B = 400$ classroom assignment permutations



This figure shows the school-level distribution of our measure for whether schools sort students into classrooms more strongly than chance would allow, given the school size, number and classroom size and student composition.

Since we do this $B = 400$ we end up with a distribution of this school concentration index based on 400 counterfactual classroom assignments for each school. We then construct S_s for each school: the share of the 400 permutations for which the actual school concentration H_s strictly exceeds the simulated concentration under random assignment H_s^{random}. For example, in a school where the actual score concentration was larger than 350 out of 400 simulated concentrations, S_s would take the value of $350/400 = 0.875$.

At this point, it is important to highlight why S_s is a superior measure of classroom sorting than H_s, especially to capture sorting on characteristics that are relatively rare. To do this, imagine trying to measure sorting based on race in a school with three classrooms and one racial minority student. Even if this school fully complies with random assignment, the measure H_s will equal 1, implying full sorting. This is because, in any classroom configuration, "all" minority students will

be in the same classroom. The measure S_s, however, will equal 0—implying perfect sorting—because in no permutation will H_s strictly exceed H_s^{random}. Generalizing based on this example, the key lesson is that S_s naturally normalizes classroom concentration to reflect the rarity of the characteristic of interest at the school level, a very useful property.[10]

Figure C1 shows the distribution of our school-level measure of classroom concentration based on test scores, S_s, for all 333 schools in the TEPS data. If all schools in TEPS would have perfectly complied with random assignment of students to classrooms, we would expect this to closely resemble a standardized uniform distribution. The figure suggests that most schools are likely complying with the random assignment mandate, yet a small but non-negligible share of schools show very high degree of sorting that is inconsistent with random assignment. Eyeballing the distribution, one could conclude that schools in the rightmost part of the distribution—with $S_s > 0.90$ which adds up to roughly 80 schools—are much more likely be defying the mandate of random assignment.

At this stage just dropping these 80 schools from our data would be rather crude. Under random assignment, we should still expect that some schools, by chance, ended up grouping students with similar test scores. Blindly trimming these schools could therefore lead to "over-trimming": removing schools that have high sorting by chance. One consequence of over-trimming is that, by removing schools that by chance ended up with sorted classrooms while leaving other schools in, it can lead to *negative* sorting tests in the trimmed sample. Over-trimmed samples do not, in themselves, lead to biased peer effect estimates. However, over-trimming could remove legitimate variation that could be crucial for identifying peer effects, resulting in a loss of power and, if peer effects are extremely non-linear, it could even introduce attenuation bias in peer effect estimates.

In the fourth and key step of our fishing algorithm, we try to disentangle schools that have strong sorting by chance from schools that are defying the mandate of random assignment using latent class models of S_s. Our preferred method is to fit a finite mixture model (FMM) of S_s to recover a

[10] A second, perhaps more subtle, lesson is that we can only interpret $P_s = 0$ as evidence of strong classroom non-sorting when the characteristic of interest is prevalent in the school (i.e., when the number of students with that characteristic exceed the number of classes in the school).

predicted probability of being a school defying the mandate of random assignment to classrooms for each school.[11] One good reason for using FMM is that, based on its estimates, we can construct the posterior probability of belonging to each latent class modeled and. Once we have identified which class is likely to capture defier schools, this gives us a direct estimate of school-level probability to be a defier, which we then use to construct our "likely defier" school flag.

For this step, there are four key choices to make: *i*) the correct model given the distribution of S_s, *ii*) the number of latent classes, *iii*) the class-level predictors (if any), and *iv*) the classification rule that flags a school as defier. We discuss these choices and our approach to making them in turn:

i. For modelling the distribution of S_s, we opted for fitting a FMM tobit to account for the censoring of S_s at 1. For other characteristics or in other datasets where S_s shows less censoring, one can always fit beta or linear regression FMMs instead. In the TEPS data all these alternatives yield similar results.

ii. We chose the number of latent classes that minimizes the Bayesian Information Criterion. In the TEPS data this was a 3-class model. Of these three classes, only one had a conspicuously large predicted mean for S_s, which was very close to 1. We identified this as the class of defier schools. The other two classes had much lower predicted means for S_s, both close to 0.5. Using the Akaike Information Criterion we would have chosen a 2-class model instead; a defier one with a predicted S_s very close to 1 and a complier one with predicted S_s close to 0.5. Both models would have classified schools near-identically. Models with more than 3 latent classes did not improve fit much but did increase optimization complexity and often had issues converging.

iii. We chose school-level class predictors that were significantly related to S_s. In the TEPS data these are schools means for: children who report being in academically gifted classrooms, parents who push for their children to be assigned to particular classrooms, ethnic minority students, private tutoring lessons before joining junior high school, and two measures of baseline student effort. All these measures were positive predictors of belonging to the defier class, most of the statistically significant at conventional levels. These predictors meaningfully improved the model performance and, since models with and without class predictors are

[11] We have also worked on procedures thar detect defier schools based on several P_s indices—to detect, for example, one type of defier school that sorts students to classrooms based on test scores, and a second type that sorts students based on their history of truancy—using unsupervised machine learning techniques such as hierarchical cluster analysis.

nested, one can make the choice to include these in the final model specification based on a likelihood ratio test.

iv. For flagging schools as defiers we constructed for each school the probability of belonging to the defier class P_s—the class with predicted S_s close to 1—based on the FMM estimates with class predictors. We then opted for classifying defier schools as schools which were *most* likely to belong to the defier class; that is, those schools for which $P_s > 0.5$. Different thresholds can of course be justified, but this is a reasonable one with a clear *a priori* justification. Our model results are not sensitive to other reasonable classification thresholds such as P_s being larger than the sum of all other predicted class probabilities.

Table C2. Summary statistics of key variables in TEPS across samples

	Mean of pre-assignment characteristics in sample:		
	TEPS	Trimmed	Estimation
Characteristic:			
Student test scores (unstandardized)	40.9	40.5	41.0
Female student	0.50	0.50	0.48
Student year of birth	1988.59	1988.59	1988.6
No. of siblings of student	1.77	1.78	1.75
Responding parent is female	0.64	0.64	0.64
Ethnic minority father	0.05	0.05	0.04
Two-parent household	0.86	0.86	0.87
Father's birth year	1958.6	1958.7	1958.6
Father has post-secondary education	0.12	0.12	0.12
Unemployed father	0.11	0.11	0.11
Household monthly income is			
NT$20,000 or less	0.11	0.11	0.10
NT$20,000-NT$50,000	0.41	0.41	0.41
NT$50,000-NT$100,000	0.35	0.35	0.35
More than NT$100,000	0.14	0.13	0.14
No. of students (approx.)	20,055	13,760	11,068

Figure 1 shows the schools eventually flagged as defiers by our fishing algorithm across the S_s distribution. We overlay the probability of being a defier school P_s (on the right y-axis) in a scatterplot, with 0.5 as a dashed horizontal reference line. Our fishing algorithm flags 106 schools where P_s exceeds 0.5 as defiers. As expected, most flagged schools have $S_s > 0.90$, though a few schools with lower values of S_s are also flagged. In the TEPS data, the algorithm failed to identify complier schools with very high S_s. It is possible, of course, that all these schools with high S_s are in fact defiers, yet it is more likely that the FMM class predictors are just not strong enough to

discern the compliers among this group. As discussed above and in Section 3.4, this could lead to over-trimming and in fact we do see some evidence of this in Table 2, which shows sorting and balancing tests in our sample trimmed of the 106 schools flagged as defiers. Yet evidence of over-trimming is not strong enough to be concerning.

As a final point in this appendix, we show that out applying our fishing algorithm in the TEPS data does not introduce any evident selectivity in our estimation samples. Table C2 shows that our initial sample including all the TEPS data remains very similar to our trimmed sample—which includes all information from schools not flagged as defiers by our fishing algorithm—, and also remains similar to the our most restricted estimation sample, which includes only students for which we observe test scores, educational inputs and other key characteristics in both TEPS waves.

Appendix D. Validating the fishing algorithm using simulated data

In this appendix, we use simulated data to validate our fishing algorithm and investigate its performance. Ideally, we would want to provide evidence from Monte-Carlo simulations of the performance of the algorithm in detecting schools that systematically sort students into classrooms. Unfortunately, we cannot provide Monte-Carlo evidence over many simulations—say, over 10,000 realizations of the same data generating process—since *i*) Steps 4 and 5 of the algorithm require making some decisions that cannot be automatized (see Box C1 in Appendix C) and *ii*) the finite mixture models in Step 4 often have convergence issues that demand making additional decisions, such as trying out different optimization procedures, grid search across different parameter values, or try out various initial latent class probabilities. Nevertheless, we provide as extensive evidence of the performance of our fishing algorithm as our setting allows, and highlight lessons learned along the way. These lessons will prove useful to researchers intending to implement our fishing algorithm in their data. In addition, we have coded flexible simulation programs in Stata which will be available with the published version of this paper.

The Data Generating Process (DGP) for our simulations

We simulate data that closely follows our empirical setting in Taiwan: students are divided into schools and, within schools, assigned to classrooms. The only characteristic that varies across students is their ability. Classrooms are simple groupings of students within schools. Students in the same classroom can end up being similar or dissimilar to one another, depending partly on chance and partly on whether their school randomly assigns students to classrooms. Schools can differ in two dimensions: whether they actively sort students of similar ability into classrooms (*sorter* schools) or not (*non-sorter* schools), and—for sorter schools—the degree to which they sort students into classrooms. In addition, we also simulate a school-level variable that predicts whether the school is sorting or non-sorting. These three parameters—the number of sorting schools, the strength of sorting within sorting schools, and the strength of the sorting school predictor—are the key parameters we vary across our simulations. All other parameters, such as school size and classroom size, are kept constant across DGPs.

Specifically, for each DGP we simulate data from 300 schools. We stochastically vary the number of students across schools between 50 and 70 with an independent uniform distribution, $U[50,70]$,

mostly as a legacy for implementing the Guryan, Kroft and Notowidgo (2009) sorting test. Their method accounted for a small negative bias in classical sorting tests by controlling for school-level leave-out-mean of student ability, but this correction only works well when there is variation in school size in the data. For our exercises, however, we implement instead the solution proposed by Jochmans (2020), who derives analytical expressions for this negative bias and proposes a bias-corrected test with better power and implementable without school-size variation. Once we have schools filled with students, we assign ability to students according to $ability \sim U[0,1]$.

At this point, we randomly determine which schools are the sorting schools that sort students into classrooms based on $ability$, and which schools are non-sorting schools. The number of sorting schools, $N_{sorting}$, is the first key parameter we vary across DGPs.

Here we also generate $predictor$, the variable predicting whether a school is a sorter or a non-sorter, given by:

$$predictor = 1\{\text{sorting school}\} * p + U[0,1] * (1 - p)$$

where $1\{\text{sorting school}\}$ is a dummy that flags sorting schools, $p \in [0,1]$ is a predictor strength parameter, and $U[0,1]$ is another independent random uniform. If p equals one, $predictor$ will be a perfect determinant of whether a school sorting students into classrooms; if p equals zero, $predictor$ will be completely uninformative. The predictor strength p is the second key parameter we vary across DGPs.

Within each school we then sort students based on the $sorting\ strength$ parameter in this school, and then sequentially assign them to similar-sized classrooms of roughly 15 students. $sorting\ strength$ is key for simulating student sorting into classrooms for some schools but not others, as is defined as:

$$sorting\ strength = \begin{cases} \theta ability + (1-\theta)U[0,1] \text{ if student is in a sorting school} \\ U[0,1] \text{ if student is not in a sorting school} \end{cases}$$

where $\theta \in [0,1]$ is the parameter that governs the sorting strength in sorting schools and we vary it across DGPs. The way this parameter works is best explained with a few examples.

When θ is one, $sorting\ strength$ equals $ability$ in sorting schools and a random uniform for non-sorting schools. This implies that in sorting schools, students will be assigned to classrooms based on their $ability$, with the first classroom having the top 15 students, the second classroom the top 15 among the remaining students, and so on. This simulates very strong sorting of students into classrooms in a scenario we refer to as "perfect stacking". In non-sorter schools, students will be randomly assigned to classrooms. If instead θ is zero, $sorting\ strength$ becomes a random uniform for all schools (sorting and non-sorting, resulting in random assignment of students to classrooms across the entire simulated data. Values of θ between zero and one will vary the strength of sorting, or stacking, in sorting schools while keeping random assignment in non-sorting schools. This θ is the second key parameter we vary across DGPs.

To make sure there is enough identifying variation in peer aggregates of $ability$, we ensure that no classroom has fewer than 10 students—which can happen because initial classroom size is set to 15 but variation in school size can occasionally lead to a classroom of fewer than 10 students. When this happens, we randomly redistribute students in these small classrooms to all other remaining classrooms, such that classrooms are always larger than 15 students.

We test the performance of our fishing algorithm using simulated data from three DGP versions that correspond to cases of particular interest for an econometrician interested in applying our method:

- $N_{sorting} = 50, \theta = 0.8, p = 0.8$: 50 strongly sorting schools with a good sorting predictor
- $N_{sorting} = 50, \theta = 0.8, p = 0.1$: 50 strong sorting schools with a weak sorting predictor
- $N_{sorting} = 300, \theta = 0.15$: all schools are weak sorters, with a good sorting predictor

The first is an ideal case where the researcher can detect the few schools that violate random assignment in the data, and has access to good enough predictors to detect whether a school is sorting systematically students. The second case showcases the limitations of our fishing algorithm when the researcher does not have access to reasonable predictors of sorter schools. The third case simulates the unfortunate situation where *all* schools sort students into classrooms, enough to invalidate random assignment in the data but with no hopes of being able to fish out defier schools with our method—or any other for that matter.

Performance of fishing algorithm

After producing data using this DGP, we then *i*) test the degree of sorting in the simulated data, *ii*) run our fishing algorithm following the steps in Box C1, *iii*) evaluate the performance of our fishing algorithm in detecting sorter schools in the simulated data, and *iv*) estimate the degree of sorting in the data once the detected sorter schools are removed. These four sets of results are presented in Panels A, B, C and D in the tables below.

We simulate five different realizations of each DGP and present the results of our fishing algorithm for each. For each simulation, we present our results in columns (1) through (5) of the tables below. The downslide of this approach is that it produces less systematic evidence of the performance of our algorithm than would Monte Carlo simulations. The upside, apart from being feasible, is that we can demonstrate the several decisions required from the researcher to use our method, explain the reasoning behind them, and showcase results of situation when, by chance, our method does not perform well.

Case 1: Few strong sorter schools and a strong class predictor

Table D1 shows the performance of the Fishing algorithm in five simulated datasets with 50 strongly sorting schools and access to a good predictor for whether schools are sorters. Panel A shows Jochmans' (2020) sorting test t-statistic estimated using the simulated student-level data. When positive and larger than critical values of the standard normal distribution, these t-statistics indicate positive sorting of students into classrooms based on ability. As expected, our simulated data shows strong evidence of sorting (first row) and this evidence is coming solely from the few sorter schools (second and third rows).

Panel B shows the steps to select the best Finite Mixture Models (FMM) to detect sorter schools. These FMMs are estimated using school-level data where the outcome is our measure of ability concentration in classrooms (S_s, see Appendix C). We first estimate FMMs with 2, 3, and 4 potential latent classes. We select the best among these models based on goodness of fit, using the smallest Bayesian Information Criteria (BIC); the BIC of the preferred model is marked in **bold** in each column. FMMs often have convergence issues—one of the reasons why we cannot produce complete Monte Carlo evidence in this Appendix. We mark models that failed to converge in

italics. After choosing the preferred number of latent classes based on the BIC, we then choose whether the preferred model will include the variable $predictor$ as a latent class predictor. For this we estimate FMMs with and without this latent class predictor and use a Likelihood Ratio (LR) test to choose between these nested models. Rejecting the null that the models are equal leads us to choose the model that includes $predictor$ as a latent class predictor. Here too, we have missing values for the p-value of this LR test when either model does not converge. Finally, we show the marginal means for each class—the average outcome predicted for schools in each latent class—in the preferred model. These correspond to the predicted level of classroom concentration in schools in each latent class. We interpret the latent class(es) with unusually high predicted means as those that identify sorter schools. These are also marked in bold.

There are three broad lessons from Panel B of Table D1. First, models with two or three latent classes are generally preferred, and models with four latent classes often have convergence issues. This relatively simple latent class structure is partly a direct result of our DGPs—which have, in fact, two latent classes of sorter and non-sorter schools—yet it confirms that the FMMs do not tend to over-fit latent classes in the data. Second, models that use latent class predictors also suffer convergence issues. This is a potential shortcoming, since we later show that these predictors can meaningfully improve the performance of our fishing algorithm. Third, there is almost always a latent class with a clearly larger predicted sorting strength, and the closer this prediction is to 1 it is that this class identifies sorter schools.

Panel C summarizes the performance of the preferred FMM for classifying defier schools—schools which, in violation of random assignment, systematically sort students into classrooms. We flag defier school as those for which the posterior latent class probability for the sorter class is larger than the sum of all the other posterior latent class probabilities, as described in Appendix C. We report four standard indicators to describe the performance of our algorithm at detecting schools that systematically sort students into classrooms: *i)* the number of schools classified as defiers (out of 300), *ii)* the percentage of schools that are correctly classified as defier schools by the fishing algorithm and are truly sorter schools, *iii)* the probability of being wrongly classified as a defier school and actually being a non-sorter school (false positives), and *iv)* the probability of being classified as a complier school and truly being a sorter school (false negative). Overall, the algorithm performs very well for this DGP: in 2 out of 5 simulations, the algorithm perfectly

separates sorter and non-sorter schools (col. (1) and col. (4)), and in 2 additional simulations it identifies no false negatives and only a few false positives (col. (3) and col. (5)).

Table D1. Fishing algorithm performance in five simulated datasets with 50 strongly sorting schools ($N_{sorting} = 50, \theta = 0.8$) and access to a good predictor for whether schools are sorters ($p = 0.8$)

Simulation number =	(1)	(2)	(3)	(4)	(5)
Panel A: Sorting t-statistic in student-level data if DGP were known					
Jochmans (2020) sorting t-statistic:					
for all schools	**6.4**	**6.6**	**6.6**	**6.9**	**6.7**
for non-sorter schools	-1.2	-0.5	-0.4	1.7	0.1
for sorter schools	**6.6**	**6.7**	**6.8**	**6.7**	**6.8**
Panel B: Finite Mixture Models on school-level data					
Model BIC for:					
2 latent classes	**316.7**	**336.7**	326.6	**313.4**	327.8
3 latent classes	327.3	344.5	**322.4**	320.6	**327.0**
4 latent classes	*318.6*	*350.5*	334.6	*325.3*	*330.3*
LR for model with sorting predictor (p-value)	0.000	-	-	0.000	-
Predicted sorting strength measure for:					
class 1	0.48	0.13	0.09	0.53	0.11
class 2	**1.02**	**0.73**	0.53	**1.01**	0.55
class 3	-	-	**1.03**	-	**1.02**
class 4	-	-	-	-	-
Panel C: Selected FMM model performance for defier classification					
Schools identified as defiers	50	225	76	50	71
Correctly classified schools	100.0%	41.7%	91.3%	100.0%	93.0%
Pr[Wrongly classified defier]	0.0%	77.8%	34.2%	0.0%	29.6%
Pr[Wrongly classified complier]	0.0%	0.0%	0.0%	0.0%	0.0%
Panel D: Sorting t-statistics in student-level data in classified schools					
Jochmans (2020) sorting t-statistic:					
for classified complier schools	-1.2	**-6.7**	**-4.2**	1.7	**-3.9**
for classified defier schools	**6.6**	**7.0**	**7.1**	**6.7**	**7.1**

In Panels A and D, numbers in bold mark values larger than the 5% critical value in the reference a standard normal distribution. In Panel B, numbers in bold mark the smallest Bayesian Information Criterion (BIC) and the largest predicted outcome mean, used to select the preferred model, and italics marks models that did not comply with convergence criteria. A missing Likelihood Ratio (LR) test p-value is missing in Panel B indicates that either the model using sorting predictors for the latent classes or the model without predictors did not converge (almost always the former).

In column (2) the fishing algorithm somewhat fails: the algorithm indicates that the majority of schools as defiers, over 50% of which are actually non-sorter schools. This failure is not complete, however, in the sense that the algorithm only becomes too stringent, but does not misclassify sorter

schools as compliant. The good news is that our exercise reveals why this failure occured: the selected FMM model in this instance could not use $predictor$ as a latent class predictor to identify the latent class with defier schools, and consequently the predicted sorting strength for this model is 0.73, well below that of all other models. The lesson for researchers applying our method here is that having access to a good predictor of whether schools are sorting will meaningfully improve the performance of our fishing algorithm, even in settings with few strongly sorting schools.

Panel D shows Jochmans' (2020) sorting test performance back in the student-level simulated data in complier schools—those classified as non-sorters by the fishing algorithm. For the two models with perfect performance (col. (1) and col. (4)), we see that the t-statistics match the non-sorter t-statistics in Panel A. For the other three models, we see negative and significant t-statistics (col. (3) and col. (5)); much more negative for the worst-performing model (col. (2)).

Negative and significant t-statistics of sorting tests become increasingly more frequent as the rate of false positives increases – that is, the probability of wrongly classifying non-sorting schools as defier schools. In Appendix C, we call this situation "over-trimming", corresponding to situations when the fishing algorithm wrongly excludes schools that are actually compliant with random assignment. The issue with over-trimming is that it could lead to censuring the distribution of peer effects.

Importantly, our algorithm can be used as a diagnostic tool for over-trimming, since a clear sign of over-trimming is a "flipping" sign of Jochmans' t-statistic: a positive and significant t-statistic in the untrimmed data (as in Panel A) and a negative and significant t-statistic in the trimmed data (as in panel D). When this occurs, we suggest going back to the FMM specification to improve the classification performance, either my changing the number of classes or by exploring additional and hopefully better class predictors. An important early sign that the algorithm is able to discern sorter from non-sorter schools is a high predicted sorting strength for at least one latent class, like in Col. (1) and (3) to (5) in Panel B.

Case 2: Few strong sorter schools and a weak class predictor

Table D2 shows the performance of our algorithm in a DGP where there are still 50 strongly sorting schools, but the researcher only has access to a much weaker predictor of whether schools are

sorters. This reflects the situation of researchers with either limited data or limited institutional knowledge to construct such predictors.

Table D2. Fishing algorithm performance in five simulated datasets with 50 strongly sorting schools ($N_{sorting} = 50, \theta = 0.8$) and but only a weak predictor for whether schools are sorters ($p = 0.1$)

Simulation number =	(1)	(2)	(3)	(4)	(5)
Panel A: Sorting t-statistic in student-level data if DGP were known					
Jochmans (2020) sorting t-statistic:					
for all schools	**6.7**	**7.0**	**6.5**	**6.7**	**6.6**
for non-sorter schools	0.6	1.6	-1.8	-0.4	-0.8
for sorter schools	**6.7**	**6.8**	**6.8**	**6.8**	**6.7**
Panel B: Finite Mixture Model selection on school-level data					
Model BIC for:					
2 latent classes	**307.5**	**322.9**	**324.1**	**308.7**	**323.4**
3 latent classes	317.2	325.4	331.5	313.3	331.4
4 latent classes	329.7	*336.9*	342.0	*330.1*	*346.1*
LR for model with sorting predictor (p-value)	0.000	0.000	0.000	0.000	-
Predicted sorting strength measure for:					
class 1	0.18	0.17	0.19	0.20	0.22
class 2	**0.73**	**0.74**	**0.78**	**0.78**	**0.77**
class 3	-	-	-	-	-
class 4	-	-	-	-	-
Panel C: Selected FMM model performance for defier classification					
Schools identified as defiers	228	233	192	199	207
Correctly classified schools	40.7%	39.0%	52.7%	50.3%	47.7%
Pr[Non-sorter school | Defier]	78.1%	78.5%	74.0%	74.9%	75.8%
Pr[Sorter school | Complier]	0.0%	0.0%	0.0%	0.0%	0.0%
Panel D: Sorting t-statistics in student-level data in classified schools					
Jochmans (2020) sorting t-statistic:					
for classified complier schools	**-5.9**	**-5.2**	**-8.8**	**-7.8**	**-6.8**
for classified defier schools	**7.1**	**7.2**	**7.1**	**7.2**	**7.0**

In Panels A and D, numbers in bold mark values larger than the 5% critical value in the reference a standard normal distribution. In Panel B, numbers in bold mark the smallest Bayesian Information Criterion (BIC) and the largest predicted outcome mean, used to select the preferred model, and italics marks models that did not comply with convergence criteria. A missing Likelihood Ratio (LR) test p-value is missing in Panel B indicates that either the model using sorting predictors for the latent classes or the model without predictors did not converge (almost always the former).

Panel A confirms that our simulated data conform to the intended DGP. Panel B illustrates that *i*) in these data the FMMs generally choose simpler 2-class structures, that *ii*) even with a much

weaker predictor the FMMs tend to prefer models with class predictors, but that *iii*) the predicted sorting strength for the high-sorting class is much weaker (between 0.73 and 0.78) than when a good class predictor is available (in Table D1). As a direct result, Panel C shows much higher rates of misclassification, driven entirely by a higher rate of non-sorter schools identified as defiers; all sorter schools are always correctly classified. As explained above, this will lead to over-trimming, Panel D confirms the presence of over-trimming: we find strong evidence of *negative* sorting in classified complier schools, and positive sorting in the classified defier schools. In sum, Table D2 corroborates the importance of having a strong sorting predictor for good performance of our fishing algorithm, but it also indicates two useful diagnostics that can tell the researcher whether the algorithm is likely to be performing poorly: a relatively low predicted sorting strength for the high-sorting latent class, and a strong flipping for the Jochmans (2020) sorting t-statistic for the classified compliers subsample. Compared to the findings of Table D1, the findings of D2 indicate that finding one or multiple strong class predictors is crucial for preventing the algorithm from over-trimming the sample.

Case 3: Weak but generalized sorting

Table D3 shows the performance of our fishing algorithm in a DGP that simulates sorting in all schools, weaker relatively to the previous DGP but strong enough that it would be detected by Jochmans (2020) t-statistic. This corresponds to setting with generalized violations of random assignment, such that no natural experiment could be salvaged from the data using our algorithm.

Panel A confirms that our simulated data conforms to this setting, producing t-statistics that significant around the 1% level. Panel B shows that *i*) the FMMs in this setting tend to choose 3- and 4-class structures, *ii*) the sorter school predictor is never statistically significant at conventional levels, which was to be expected since all schools are sorters, and *iii*) the predicted sorting strength in the high-sorting latent class is higher than in Table D2 but lower than in Table D3. This high predicted sorting strength results in relatively few schools identified as defiers, as show in Panel C. Because the FMMs classify as defiers the schools where the strongest sorting occurs, Panel D again shows strong flipping in the Jochmans (2020) t-statistic.

Overall, Table D3 indicates that situations where all schools sort students into classrooms (generalized sorting) compared to clustered sorting (cases 1 and 2) are characterized by *i)* relatively

complex latent class structures, *ii)* relatively low model fit yet *iii)* high predicted sorting strengths for the high-sorting latent class even in the absence of good sorting school predictors (Panel B), and *iv)* flipping of the Jochmans (2020) sorting t-statistic for identified complier schools (Panel D).

Table D3. Fishing algorithm performance in five simulated datasets with all weakly sorting schools ($N_{sorting} = 300, \theta = 0.15$)

Simulation number =	(1)	(2)	(3)	(4)	(5)
Panel A: Sorting t-statistic in student-level data if DGP were known					
Jochmans (2020) sorting t-statistic:					
for all schools	**3.6**	**4.5**	**4.3**	**4.6**	**3.0**
for non-sorter schools	-	-	-	-	-
for sorter schools	**3.6**	**4.5**	**4.3**	**4.6**	**3.0**
Panel B: Finite Mixture Model selection on school-level data					
Model BIC for:					
2 latent classes	105.0	103.7	85.4	104.9	99.4
3 latent classes	**93.4**	92.5	**66.3**	**103.1**	95.0
4 latent classes	95.1	**84.9**	70.4	*88.7*	**91.9**
LR for model with sorting predictor (p-value)	0.818	0.280	0.170	0.066	0.850
Predicted sorting strength measure for:					
class 1	0.16	0.15	0.09	0.13	0.05
class 2	0.56	0.41	0.52	0.52	0.38
class 3	**0.93**	0.72	**0.92**	**0.90**	0.81
class 4	-	**0.95**	-	-	**0.96**
Panel C: Selected FMM model performance for defier classification					
Schools identified as defiers	79	66	82	107	46
Correctly classified schools	26.3%	22.0%	27.3%	35.7%	15.3%
Pr[Non-sorter school | Defier]	-	-	-	-	-
Pr[Sorter school | Complier]	-	-	-	-	-
Panel D: Sorting t-statistics in student-level data in classified schools					
Jochmans (2020) sorting t-statistic:					
for classified complier schools	**-4.7**	**-3.6**	**-3.9**	**-4.7**	**-3.2**
for classified defier schools	**6.7**	**6.9**	**7.5**	**7.5**	**5.6**

In Panels A and D, numbers in bold mark values larger than the 5% critical value in the reference a standard normal distribution. In Panel B, numbers in bold mark the smallest Bayesian Information Criterion (BIC) and the largest predicted outcome mean, used to select the preferred model, and italics marks models that did not comply with convergence criteria. A missing Likelihood Ratio (LR) test p-value is missing in Panel B indicates that either the model using sorting predictors for the latent classes or the model without predictors did not converge (almost always the former).

A practitioner's guide for researchers wanting to use our fishing algorithm

Our fishing algorithm combines several intuitive steps which are nonetheless somewhat technically complex. Drawing on the lessons illustrated in this section and on our own experience in developing this algorithm, we make the following suggestions to researchers intending to use our method:

1. **Strive to find predictors of whether a school sorts students into classroom, even if these predictors are not perfect**. Good predictors will meaningfully improve the performance of our method, even if it can still be applied without them. Place more trust in applications with institutionally sound sorting predictors that are also statistically and quantitatively strong inputs in your latent class model.
2. **Your latent class that captures sorting schools will have a predicted sorting strength close to or exceeding 1**. By the nature of our measure of sorting strength, sorting schools should have strengths very close to or greater than 1. Latent classes with predicted sorting strengths much below 1 are therefore more likely to also capture non-sorting schools, increasing over-trimming problems. If your latent class model is not identifying classes with high enough predicted sorting strengths, this could be a sign that *i*) the class structure is not complex enough (solved by testing models with more latent classes), *ii*) your sorting school predictors are not good enough (solved by finding better predictors or a better structure for existing ones), or *iii*) sorting is too widespread in your data (only solved, sadly, by finding other data that reflects a better natural experiment).
3. **Beware of sorting test flipping**. Sorting test flipping—a large and positive sorting t-statistic in the whole data and a large and negative sorting t-statistic in the subsample of identified complier schools—is a sign of either over-trimming or widespread sorting.


POSTERA CRESCAM LAUDE
THE UNIVERSITY OF
MELBOURNE
MELBOURNE
INSTITUTE