Detecting Local Non-Compliance with Random Treatment Assignment in Quasi-Experiments, with an Application to Ability Peer Effects

Alexandra de Gendre Nicolás Salamanca*

September 2024

Abstract

We develop a method to detect local non-compliance with a random treatment to recover valid quasi-experiments from existing data. Our approach combines simulation-based methods and latent class modelling in an intuitive way, can detect and characterize non-compliant risk sets, and is computationally undemanding. To illustrate its usefulness, we use it to estimate ability peer effects in Taiwan, where we have unusually rich data and a national mandate to randomly assign students to classrooms within schools, but with partial non-compliance. After recovering a valid quasi-experiment, we first estimate ability peer effects in line with other studies. We then document precisely estimated null effects on 18 potential mechanisms, many of which have been hypothesized but never tested. Our application shows how to use our method to expand the causal evidence base using existing datasets.

JEL: I23, I26, D13

Keywords: treatment assignment non-compliance, random assignment, peer effects, test scores, educational investments

*de Gendre: Corresponding author. Department of Economics, The University of Melbourne; IZA and ARC Centre of Excellence for Children and Families over the Life Course (LCC). a.degendre@unimelb.edu.au. Salamanca: Melbourne Institute: Applied Economic & Social Research, The University of Melbourne; IZA and ARC Centre of Excellence for Children and Families over the Life Course (LCC). n.salamanca@unimelb.edu.au. An earlier version of this paper circulated under the title "On the Mechanisms of Ability Peer Effects." We thank Xavier D'Haultfoeuille, Brian Graham, Sami Stouli, Chris Karbownik, David Figlio, Olivier Marie, and especially Jan Feld and Ulf Zölitz for their helpful comments. Boer Xia provided excellent research assistance for this project. Data analyzed in this paper were collected by the research project "Taiwan Education Panel Survey: The First Wave", sponsored by Academia Sinica, Ministry of Education, National Academy for Educational Research and National Science Council. License number: 012018005. The Survey Research Data Archive, Academia Sinica is responsible for the data distribution. The authors appreciate the assistance in providing data by the institutes and individuals aforementioned. We 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 Government through the Australian Research Council's Centre of Excellence for Children and Families over the Life Course (Project ID CE140100027 and CE200100025). The views expressed herein are the authors' own.

1 Introduction

The key hurdle for producing high-quality empirical evidence is pairing the right experiment with the right data. Experiments and quasi-experiments are both plagued by problems such as imperfect comparison groups, manipulation of treatment assignment, spillovers across treatment groups, and selective attrition. In the rare settings where a great experiment is available, high-quality data seldom are. Secondary data often provide limited power, outcomes are mismeasured or unmeasured, and covariates are not rich enough to produce convincing evidence of a valid quasi-experiment. And while fit-for-purpose datasets—such as the ones explicitly collected for an experiment or targeting a specific quasi-experiment—are becoming more common, those can often only collect short-term measures on surrogate outcomes.

In this paper we develop a widely applicable tool to allow pairings of good experiments and good data that would otherwise not be available. Our method is a statistical procedure that allows researchers to recover a valid quasi-experiment from a setting where imperfect experimental variation is available. By imperfect, we mean where the random variation from an experiment or quasi-experiment failed in some risk sets—groups of observations facing the same risk of treatment—to such an extent that the overall experiment would be deemed invalid. This issue is common in quasi-experiments that exploit idiosyncratic variation in random matching, such as student or teacher assignment to classes within schools, patient assignment to doctors within hospitals, or case assignment to judges within courtrooms. More broadly, this issue can occur in settings where there are institutional reasons to manipulate a quasi-randomly assigned policy, such as the disbursement of subsidies by municipal authorities, the award of scholarships to marginal applicants by panels, or the offer of available spots to prospective students by oversubscribed schools. Our tool combines permutation-based methods and latent class models to identify and remove risk sets where random assignment to treatment failed. Researchers can then analyze the remaining data with standard methods. Since our tool "fishes" out risk sets where randomization was not complied with, we call it the Fishing Algorithm.

In the first part of the paper, we describe the Fishing Algorithm in detail. The intuition is simple. We first construct a statistic that measures whether there is evidence of random treatment assignment within risk sets. An example could be the t-statistic from a balancing test that regresses predetermined characteristics or placebo outcomes on the treatment status. Next, we use permutation methods to recover the distribution of this statistic under the null hypothesis of random assignment to treatment for each risk set. By comparing the actual statistic to the distribution of simulated statistics under the null of random assignment, we can assess whether treatment assignment in any one risk set looks too extreme to have happened by chance. This procedure is similar to constructing randomization inference p-values, but applied to each risk set individually. It may be tempting to classify all risk sets with an empirical value approaching one as cheaters, but this would lead to wrongly excluding some risk sets that, by pure chance, look inconsistent with random assignment. Instead, in our final step we use data at the risk set level and latent class models to (1) identify a latent class of risk sets that are not compliant with random treatment assignment, (2) characterize these risk sets in terms of their observable characteristics, and (3) recover the posterior probability that any one risk set belongs to the non-complier class. With these results at hand, we can then classify all risk sets as compliant or non-compliant with random assignment, remove

the non-compliant ones from our data, and analyze the remaining data as if the quasi-experiment in it was successful. Using simulated data, we show that the Fishing Algorithm is remarkably effective for identifying non-compliant risk sets.

In the second part of the paper, we illustrate the value of the Fishing Algorithm by using it to estimate ability peer effects and their mechanisms using Taiwanese data. This context is ideal for three reasons. First, we have access to the Taiwan Educational Panel Survey (TEPS), an exceptionally rich longitudinal survey of students within classrooms in schools including measures of achievement and many unexplored mechanisms. Second, there is a national mandate in Taiwan to randomly assign students to classroom within schools which should have generated an excellent quasi-experiment to answer these questions. Yet third, there seems to have been meaningful non-compliance with the random assignment mandate in some schools, our risk sets in this application, which we identify and solve using our method. We focus on the linear-in-means model, which has strong behavioral micro-foundations (Ushchev and Zenou, 2020; Boucher et al., 2024) and is the most used peer effect model. Boucher et al. (2024) show that this model is appropriate for outcomes such as GPA, joining social clubs, self-esteem, and exercise.

Standard tests produce definitive evidence that the random assignment mandate does not hold in the entire TEPS data. Our method then identifies 102 out of the 333 schools in our data that sort students into classrooms based on ability in non-compliance with the national random assignment mandate. These sorter schools are more likely to offer academically or artistically gifted classes, according to their students, which matches circumstantial evidence suggesting that schools use off-the-books gifted classrooms as a tool to bypass random assignment mandate. After trimming these schools from our data, we no longer find evidence of non-compliance with random assignment.

We then use this trimmed dataset to first show that in Taiwan having peers with one standard deviation higher average test scores increases students' own test scores by 3.7 percent of a standard deviation. These estimates are in line with the literature on ability peer effects in other settings and levels of education with randomly assigned peers (e.g., Feld and Zölitz, 2017). We then add to this literature by estimating the impact of higher-achieving peers on 18 potential mechanisms for these effects, many of which have not been explored before. In particular, we are the first to be able to jointly observe responses of students, parents, and teachers to exposure to high-ability peers. We find precisely estimated null effects on most of these. We do find some evidence that higher-achieving peers increase parental time investment in their children and decrease student's ratings of the quality of the school environment.

We finish the paper by providing a few additional results on the effects of studying with higher ability peers. We also perform an extensive set of sensitivity analyses that demonstrate the reliability of the Fishing Algorithm to recover a valid quasi-experiment and produce our target achievement peer effect estimates. We 1) conduct additional tests for conditional random assignment of students to classrooms, 2) assess result sensitivity to key tuning parameters in our Fishing Algorithm, 3) implement an alternative solution that uses some of the Fishing Algorithm's output for re-weighting the entire dataset, 4) assess sensitivity to alternative measures of student and peer ability, 5) implement corrections for incomplete sampling of classrooms, 6) adjust our inference using randomization inference and multiple hypotheses testing adjustments, and 7) conduct an extensive exploration of effect heterogeneity. In addition, we also provide suggestive evidence on 8) the long-term effects up to age 24, 9) the degree to which our tested

mechanisms mediate achievement peer effects in these data, and 10) the potential for higher-ability peers to change the technology of skill formation.

Our key methodological contribution, the Fishing Algorithm, can solve failure-to-randomize problems that occur in many different ways, and can detect non-compliance with random treatment assignment at the risk set level in datasets where actual treatment is observed but not treatment assignment. This setting is common in many quasi-experiments that exploit idiosyncratic variations in matching, such as teacher assignment to students, patient assignment to doctors, or case assignment to judges. In such settings with partial compliance to random assignment and *a priori* no reliable way to know where compliance occurs, researchers often try to account for systematic non-compliance with random assignment by controlling for additional characteristics beyond balancing controls. This approach complicates the interpretation of estimates and can only recover causal estimates under strong conditional ignorability assumptions. The Fishing Algorithm, instead, focuses on detecting entire non-compliant risk sets. It thus offers a transparent alternative to improve the validity of quasi-experimental research designs based on conditional random assignment, which does not require conditioning on pre-assignment covariates.

The marginal contribution of our paper in terms of method and application is best understood within the literature on forensic economics. Most studies in this literature focus on detecting the incidence of cheating (see e.g., Le Moglie and Sorrenti, 2022; Ghanem and Zhang, 2014), yet the Fishing Algorithm allows researchers to also study *who* cheats. Methodologically, our approach shares most with that of Chalendard et al. (2023) to detect corruption in customs. Our Fishing Algorithm can be seen as a generalization and extension of their approach to detect manipulation of inspector assignment to custom declarations in Madagascar's main port. We generalize by explicitly describing the settings, steps and conditions where this simulation-based approach can be used to detect non-compliance with random treatment assignment in specific risk sets. We extend by embedding in our method a way to jointly classify and characterize non-compliant risk sets using finite mixture models. In that sense, our Fishing Algorithm combines both their simulation-based and model-based approaches to detect corruption in customs, albeit without requiring Bayesian Markov Chain Monte Carlo methods for estimation. This also permits a broader application of our method to other settings. In addition, our method requires little institutional knowledge to detect risk sets where cheating occurs (also unlike e.g., Oliva (2015) or Jacob and Levitt (2003)). In fact, rather than requiring specific institutional information on who cheats to work well, the Fishing Algorithm can instead provide information on who is likely cheating. This information can be contrasted with institutional knowledge to verify whether the algorithm is working sensibly. In spirit and in our application, our paper also shares a lot with Oliva (2015), which first identifies cheater autocenters and then uses data on non-cheating autocenters to estimate the predicted true emissions test pass rate per auto. Similarly, we use the Fishing Algorithm to identify schools that do not comply with the mandate of random assignment in Taiwan, and use data on compliant schools to estimate academic peer effects in a new setting and conduct an extensive study on their potential mechanisms.

Finally, the findings of our application contribute to an extensive literature on achievement peer effects.[1]

1. The past two decades have yielded an impressive number of excellent empirical studies on achievement peer effects (e.g., Sacerdote, 2001; Zimmerman, 2003; Hanushek et al., 2003; Hoxby and Weingarth, 2005; Vigdor and Nechyba, 2007; Kang et al., 2007; Ammermueller and Pischke, 2009; Duflo, Dupas and Kremer, 2011; Lavy, Silva and Weinhardt, 2012; Burke and Sass, 2013; Carrell, Sacerdote and West, 2013; Feld and Zölitz, 2017; Booij, Leuven and Oosterbeek, 2017). For brevity, we focus on

We produce causal estimates of exposure to higher achieving peers in Taiwan, a new setting, and show that they are commensurable to what others have found elsewhere. We then conduct the largest exploration of the many potential mechanisms behind achievement peer effects within a single setting.[2]. Some of the effects we estimate have been explored before (see e.g., Feld and Zölitz (2017) on perceived quality of peer interactions and Bursztyn and Jensen (2015) and Bursztyn, Egorov and Jensen (2019) on social pressure and effort provision). Yet most of our estimates are new to the literature, in particular those on parents' and teachers' behavioral responses to classroom ability composition. In fact, no study before has been able to test as many candidate mechanisms as we do, let alone in one single quasi-experimental setting. Our findings rule out a host of mechanisms hypothesized in this extensive literature, and should result in a large shift in prior beliefs and therefore be very informative (Abadie, 2020).

2 Non-Compliance with Random Assignment and The Fishing Algorithm

In this Section we explain the steps of our Fishing Algorithm. The algorithm detects local non-compliance with random treatment in datasets where actual treatment is observed but not treatment assignment. This setting is common in many quasi-experiments that exploit idiosyncratic variation in matching, such as peer or teacher assignment to students, patient assignment to doctors, or case assignment to judges. Our algorithm is also useful in settings such as the randomized nation-wide evaluation of a program where implementation is done at the municipal level and some of these municipalities might have tampered with the random component of the evaluation. As an example, in 2022 this sort of behavior took place in the GiveDirectly cash transfer program in the Democratic Republic of Congo. This could have compromised an evaluation of the program should it had been taking place at the time.

2.1 Detecting Overall Non-compliance with Random Treatment

Assume data $\mathscr{D}$ has been generated from an experiment that intends to randomly assign a treatment to unit $n = 1, \cdots, N$ in treatment group $c = 1, \ldots, C$ within risk set $r = 1, \ldots, R$. The data on each unit is $\mathbf{D}_{ncr} = (\mathbf{X}'_{ncr}, T_{ncr})$ where $\mathbf{X}_{ncr}$ is a vector of characteristics (including pre-determined at treatment and outcomes) and T_{ncr} is the treatment status. Treatment groups are groups of units where the treatment is assigned, in the sense that the treatment status of unit n will only depend on information of units in

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 et al., 2014) and beliefs (Boisjoly et al., 2006). There is a consensus that achievement peer effects are generally positive but small, and that the size of these effects can depend non-linearly on students' own achievement.

2. Recent studies with high-quality experiments and quasi-experiments that do explore mechanisms argue that achievement peer effects could be largely driven by three types of mechanisms: i) student effort provision (e.g., Kang et al., 2007; Brunello, De Paola and Scoppa, 2010; Feld and Zölitz, 2017; Xu, Zhang and Zhou, 2020), ii) group dynamics and network formation (e.g., Lavy and Schlosser, 2011; Lavy, Paserman and Schlosser, 2012; Bursztyn and Jensen, 2015; Brady, Insler and Rahman, 2017; Feld and Zölitz, 2017; Zárate, Forth.; Carrell, Sacerdote and West, 2013; Booij, Leuven and Oosterbeek, 2017), and iii) teacher effort or school resources (e.g., Duflo, Dupas and Kremer, 2011; Chetty et al., 2011; Hoekstra, Mouganie and Wang, 2018; Lavy, Paserman and Schlosser, 2012; Booij, Leuven and Oosterbeek, 2017; Feld and Zölitz, 2017; Aucejo et al., 2020)

their treatment group, such that $T_{ncr} = f(\{\mathbf{D}_{ncr}\}_{n \in c})$.[3] Risk sets, in borrowed terminology from Angrist, Hull and Walters (2022), are sets of treatment groups where treatment is as good as randomly assigned. Differences in the treatment assignment risk of units are possible across risk sets but not within them. To fix ideas, consider the assignment of cars to test lines within auto centers in Mexico, as in Oliva (2015), or the assignment of students to classrooms within schools in Taiwan as in our application in Section 3.

The first step in our process is to assess whether the experiment that generated $\mathscr{D}$ was well executed, in the sense that it randomly assigned T_{ncr} conditional on risk set r. Various tests have been proposed for this purpose, yet all of them involve calculating a statistic (or series of statistics) $B = B(\mathbf{P}'_{ncr}, T_{ncr})$ where $\mathbf{P}_{ncr} \subset \mathbf{X}_{ncr}$ is a subset of characteristics pre-determined with respect to treatment T_{ncr}. The usual procedure tests the null hypothesis that T_{ncr} is independent of $\mathbf{P}_{ncr}$ conditional on the risk set, and rejects this hypothesis if B is larger than some known critical value. In most applied microeconomics articles this will take the form of a balancing test, where B is a t-statistic to be compared to the critical values of a standard t distribution. The typical article will present several such statistics to be judged separately. Multiple tests can also be combined into a single one that produces a single F-statistic in the spirit of the left-hand side test of Pei, Pischke and Schwandt (2019), where B is then compared to the corresponding critical values of the F distribution.

Our Fishing Algorithm becomes useful when there is evidence that the experiment generating data $\mathscr{D}$ was not well executed, in the sense that T_{ncr} does not seem random conditional on risk set r.

2.2 The Fishing Algorithm

The Fishing Algorithm is a data-driven method to detect entire risk sets where the experiment generating $\mathscr{D}$ was not successful. Once these risk sets are detected, we can remove them from the data to keep and analyze $\mathscr{T} \subset \mathscr{D}$ where the experiment was well executed. We refer to $\mathscr{T}$ as the trimmed data, since it is produced by trimming risk sets that are likely non-compliant with the intended experiment.

Our algorithm combines risk-set-specific statistics that capture contribution of each risk set to B, permutation-based constructions of these statistics, and an explicit classification of $\mathscr{D}$ into subsets that are compliant and non-compliant with the intended experiment using a latent-class model. The intuition behind the procedure is simple and its implementation is fast. Assuming you have already identified the test statistic B that suggests the experiment generating $\mathscr{D}$ was not well executed, our algorithm steps are:

3. Grouped treatment of this form includes situations where whole groups share the same treatment (e.g., county-level data with a state-level treatment shared across counties) or where the treatment status for unit n only depends on the treatment status of all other units in its group (e.g., student-level data where networks are formed at the classroom level).

The Fishing Algorithm

1: Construct risk-set specific measures $b_r = b_r(\mathbf{P}'_{ncr}, T_{ncr}) \; \forall \; r \in \mathscr{D}$ proportional to the contribution of risk set r to the statistic B. These measures do *not* need to be a direct construction of B with data from risk set r alone, but the condition $\frac{\partial B}{\partial b_r} > 0$ should hold for all r at a minimum.

2: For each risk set r, construct K alternative assignments of treatment using permutations, such that each T^k_{ncr} for $k = 1, \ldots, K$ is consistent with a well-executed version of the intended experiment generating $\mathscr{D}$ (i.e., such that T^k_{ncr} is independent of $\mathbf{P}_{ncr}$ conditional on the risk set). For each constructed alternative assignment, construct also $b^k_r = b^k_r(\mathbf{P}'_{ncr}, T^k_{ncr})$. Finally, for each risk set construct the measure $S_r = K^{-1} \sum_{k=1}^{K} \mathbb{1}[b_r > b^k_r]$ where $\mathbb{1}[\,\cdot\,]$ is an indicator function that takes the value of one if its argument is true and zero otherwise.

3: Estimate the latent probability that each risk set is compliant with random treatment assignment using the Finite Mixture Model (FMM) $\mathscr{L}(\theta) = \prod_{r=1}^{K} \sum_{i=1}^{L} \pi_i f_i(S_r)^{c_{ri}}$ where $\pi_i = \frac{exp(\mathbf{Z_r}'\psi_i)}{\sum_j exp(\mathbf{Z_r}'\psi_j)}$ is the probability for the i^{th} latent class, $\mathbf{Z_r}$ is an optional vector of risk-set level variables included in the model as latent class predictors, $f_i(\cdot)$ is the conditional density function for S_r in the i^{th} latent class, c_{ri} is a binary variable that indicates whether risk set r belongs to latent class i, and $\theta = \{\psi_i, \varphi_i\}_{i=1}^{L}$ collects all main and ancillary class-specific parameters. The number of latent classes can be determined via information criteria (e.g., Bayesian or Akaike). The latent class of non-compliant risk sets $l \in \{1, \ldots, L\}$ is defined as the one with the highest posterior predicted mean of S_r, such that $\hat{\mu}_l > \hat{\mu}_i \; \forall \; i \neq l$, where $\hat{\mu}_i = K^{-1} \sum_{r=1}^{K} f_i(S_r; \hat{\theta})^{c_{ri}}$ is the predicted mean of S_r for latent class $i = 1, \ldots, L$. $\hat{\mu}_l$ should be close to 1.

4: Define your trimmed data $\mathscr{T}$ as the subset of risk sets for which $\tilde{\pi}_{lr} < \tau$ where $\tilde{\pi}_{lr} = \frac{\hat{\pi}_l f_l(S_r)}{\sum_{i=1}^{L} \hat{\pi}_i f_i(S_r)}$ is the predicted posterior probability of risk set r belonging to the latent class of non-compliant risk sets l. The trimmed data is therefore the subset of risk sets for which this predicted posterior probability falls below the predetermined threshold τ, with $\tau \approx 0.5$ as a natural candidate. Finally, use the trimmed data $\mathscr{T}$ to analyze the intended experiment.

There are several things to note in each step of the algorithm. In Step 1, using more sensitive measures of b_r (i.e., measures that increase more sharply when the data from risk set r is not consistent with the experimental treatment assignment) will yield better results. Non-linear statistics as well as statistics with exponential penalty terms are natural candidates. Also, while the statistic B should account for sampling variation (e.g., t-ratios that normalize partial correlations between $\mathbf{P}_{ncr}$ and T_{ncr} by their standard error), it might be ill-advised to account for sampling variation at the risk set level in b_r. The reason is that trying to account for the sampling variation within each risk set, which might be substantial, can decrease the sensitivity of b_r.

In Step 2 there are three important insights. First, for each risk set r, S_r measures how extreme is the realization of b_r relative to K counterfactual b^k_r which are, by design, constructed from a Data Generating Process (DGP) that respects the intended experiment that ought to have produced $\mathscr{D}$. Second, the procedure keeps the composition of risk sets and treatment groups intact across permutations, and respects the intended treatment structure as defined within treatment groups. For example, if in some risk sets the probability of treatment is higher than in others, the procedure will respect and incorporate that.

And third, S_r shares many similarities with empirical p-values. In particular, $S_r \in [0,1]$ by construction and, if $\mathscr{D}$ were indeed produced by a well-executed experiment, we should expect S_r to be uniformly distributed across risk sets within its range. The more inconsistent data in risk set k is with a well-executed experiment, the closer S_r will be to one. With enough ill-behaved risk sets, the distribution of S_r will become left-skewed, and when the evidence on sufficiently many of these ill-behaved risk sets is strong enough, the distribution of S_r will show a probability mass at one. Probability masses at zero indicate forced balancing, such as the one generated by the Cube Method (see e.g., Davezies, Hollard and Merino, 2024).

In Step 3, a Finite Mixture Model (FMM) $\mathscr{L}(\cdot)$ that respects and uses the structure of S_r will generally perform better in classifying the data. For example, since S_r is censored below at 0 and above at 1, it is natural to use finite mixture Tobit regressions. In instances where the distribution of S_r shows very large masses at zero or one, modelling these mass points separately and explicitly—such as with zero-inflated models —can bring additional gains. However, Maximum Likelihood estimation of these models is known to have convergence issues. Aside from that, one can also enrich the model with risk-set-specific covariates derived from the data D_{ncr} (e.g., the mean or median of an individual-level variable for each risk set r). These covariates can improve the model fit, further differentiate the latent classes, and their coefficients (estimated as part of the set $\hat{\psi}_l$) can be used as ancillary statistics for characterizing l, the latent class of non-compliant risk sets.

Finally, in Step 4 the choice of τ can be consequential. A natural choice is to make it so that the trimmed data follows a "most likely not in l" rule: $\tau = \sum_{c \notin l} \hat{\pi}_c$. Other rules are possible, and will generally result in more severe trimming of the data. The posterior probability estimates $\hat{\pi}_l$ can also be used as ex-post weights to improve efficiency.

Our Fishing Algorithm is broadly applicable. Our derivation of the general principle in terms of risk sets and general test statistics accommodates most (quasi-)experiments. We have also explored adaptations to detect multidimensional non-compliance (e.g., where treatment assignment in one risk set is correlated with one characteristics whereas in another risk set treatment assignment is correlated with a different characteristic). Note also that, in general, our Fishing Algorithm is not equivalent to controlling for observable characteristics to achieve conditional balancing. Our approach combines knowledge of the intended level of treatment assignment and the nature of the treatment to identify risk sets where random treatment non-compliance occurs. Once non-compliance is detected, we entirely remove these risk sets from the data rather than trying to keep them and account for the non-randomness via controls. Only very stringent selection on observable procedures should be able to capture endogeneity as we do, and even then these would have to apply flexible control functions at the cost of many degrees of freedom and interpretability of estimates (and that is assuming that rich enough controls to fully capture selection in non-compliant risk sets are available).

In the remainder of the paper we explore its application to estimating ability peer effects in Taiwan. In Appendix A we provide validation of the algorithm using simulated data following the Taiwanese structure described below.

3 An Application to Ability Peer Effects in Taiwan

We illustrate the use of our Fishing Algorithm to estimate the effects and potential mechanisms of studying with higher ability peers in Taiwan. As we describe below, the Taiwanese context is ideal for this application since (1) we have exceptionally rich longitudinal data of students within classrooms in schools including measures of ability and mechanisms and (2) there is a national mandate to randomly assign students to classroom within schools which should have generated an excellent quasi-experiment, but (3) several schools seem to not comply with the random assignment mandate. We identify and exclude these schools from further analyses using our method.

3.1 Junior High Schools in Taiwan

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

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., 4^{th} out of 72 countries in PISA 2015; Law (2004)).

Crucial in our application, since at least the 1990s the government requires junior high schools to randomly assign students to classrooms. This assignment of students to classrooms within schools is often referred to as "mixed-ability class grouping" or grouping for short.[4] Junior high schools are required to implement and maintain grouping in all grades. Grouping is conducted by the municipality, city, or county government, by a designated school throughout, or by each school when given approval to handle grouping itself. Grouping may be done by "giving the students a test then forming each class using an S-sequence listing of all the test scores, a public drawing of lots, or using computer random number generation." The S-sequence listing assignment refers to sorting all students in a school by test

4. In 2001 the random assignment of junior high school students and homeroom teachers (Dao Shi) to classrooms within schools was regulated by Article 12 of the *Primary and Junior High School Act*. This act was later superseded by the *Regulations Governing Mixed-ability Class Grouping and Formation of Learning Groups in Elementary and Junior High Schools*, which were drawn in accordance with its provisions. Currently, these regulations can be found in the *Primary and Junior High School Act*. Its legislative history traces back to the Presidential Decree (68) Tai-Tong(Yi)-Yi-Zi No. 2523 promulgated on May 23, 1979, which set out the initial 22 articles regulating pre-secondary schooling. Since then, the Act has been amended 16 times, yet to the best of our knowledge none of the amendments altered the original principles of random assignment to classrooms. There are few exceptions for these grouping regulations, yet none of them are relevant for this study. One exception, introduced later, is for schools in areas with high demand for foreign professionals and outstanding experts in the field of science and technology from outside Taiwan, which allows schools to set up special programs for their children. Another exception, described in the *Education Act for Indigenous Peoples*, allows schools with a minimum share of Indigenous students to implement a single-track system in senior high school, but makes no differential provisions for junior high school, where there is no tracking.

performance and then sequentially assigning the best-performing student to Classroom 1, the second-best performing student to Classroom 2, and so on, returning to Classroom 1 after a student has been assigned to each available classroom. This grouping results in classrooms that are more balanced and heterogeneous in terms of ability than predicted by chance. Public drawing of lots and random number generator assignment should directly result in random assignment to classrooms within schools. Within seven days of students being assigned to classrooms, the school assigns homeroom teachers (Dao Shi) to each class by public drawing of lots in a procedure invigilated by representatives of the teacher and the parents associations. The implementation of grouping by the guidelines below are major items to be referred to when conducting school evaluation and performance assessment and selection of principals. Violations are immediately dealt with.[5]

At the end of junior high school, students take the National Basic Competence Test, which plays 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 time and effort preparing for these exams, and schools regularly organize practice exams and other forms of preparation. Parents are also engaged 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.

Our application combines the Taiwanese natural experiment with uniquely rich survey data from the Taiwanese Education Panel Survey (TEPS) to estimate ability peer effects and 18 potential mechanisms.

3.2 The Taiwan Education Panel Survey

The TEPS is a project jointly funded by the Ministry of Education, the National Science Council, and the Academia Sinica. It is a nationally representative longitudinal survey of the education system in junior high school, senior high school, vocational senior high school, and junior college. It is a multiple respondent survey, collecting linked information on students, parents, teachers, and school administrators.

The junior high school sample of the TEPS allows us to measure student ability right after assignment to classrooms at the beginning of junior high school. The 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 the very beginning of students' first year of junior high school. The second wave was collected in 2003, at the beginning of the students' last year of junior high school.[6]

5. There is strict invigilation of this procedure, including the formation of "classroom grouping promotion committees" and term of reference for allowed classroom extracurricular activities. Schools that have been approved to handle their grouping themselves are subject to the same regulation, need to publicly announce the public grouping process date and time, invite parents as observers, and have the process supervised by government personnel. Incoming or transfer students are assigned to classrooms based on drawing of lots or randomly generated numbers. If the number of classrooms is changed over time, the new classrooms should also be generated by the same procedure.

6. Further data on some of these respondents was collected in waves three and four (11^{th} and 12^{th} grades) of the initial TEPS project and in four additional waves (around ages 20, 24, 25 and 30) of the TEPS Beyond project. However, only a small sample of the original respondents was contacted for these waves and this sample is highly selective. Noting these limitations, in Section 3.5 we discuss results on longer-run outcomes using data from these later waves.

Three key data features of TEPS 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 333 randomly selected junior high schools were sampled (45 percent of all high schools in the country at the time), with 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 was then randomly sampled. In each of these classrooms, around 15 students were then randomly sampled, amounting generally to a random half of the classroom, since the mandated maximum class size at the time was 35 students per class. This sampling framework is similar to that of the National Longitudinal Study of Adolescent to Adult Health (Add Health), a panel study of middle and high school pupils in the United States.[7]

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 students' cognitive ability and analytical reasoning, and was specifically designed to capture gradual learning over time. The test contains 75 multiple-choice questions, covering general reasoning, mathematics, Chinese and English. These questions were taken from an extensive bank of questions which includes adapted questions from international standardized tests, as well as questions provided by education and field experts in Taiwan. The Comprehensive Analytical Ability test score, constructed as the sum of all correct answers, provide excellent measures of academic ability for students and their peers. Importantly, the tests were administered by the TEPS research team, they were externally graded and the scores were not disclosed to either students, parents, teachers or school administrators.

Third, TEPS provides a wealth of questions measuring student behaviors, attitudes and beliefs in and outside the school environment, parent-child interactions and parental investments, as well as detailed information on teachers and school administrators. Many of these measures have multiple raters, combining questions asked to students, parents, teachers and school administrators. We aggregate these questions to construct an extensive battery of measures of student, teacher and parent inputs in the education production function. This large set of input measures allows us to comprehensively explore potential mechanisms behind academic peer effects.

We identified key inputs of students, parents and teachers in the education production function that are potential mechanisms of ability peer effects based on previous literature. Table B.2 provides a high-level summary of the measures of academic ability and educational inputs which we construct using the TEPS data, listing for each measure the number of items used and the number of unique values each measure takes.

For inputs with multiple potential measures, we first identify entire blocks of items in the questionnaires—e.g., blocks of items related to study effort reported by students, parents and teachers. We eliminate very low correlates to maximize the informational content of each index and reduce noise. To do this, we compute Spearman correlations between all items under consideration, assess their inter-item consistency,

7. 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 and Isphording, 2017; Card and Giuliano, 2013; Bifulco, Fletcher and Ross, 2011; Calvó-Armengol, Patacchini and Zenou, 2009).

and perform an exploratory factor analysis. Once we narrow down the list of items for a scale, we perform an additional confirmatory factor analysis to validate these items and ensure their factor loadings have similar magnitudes. Finally, for each of these potential mechanisms and in each wave, we construct a summative scale that adds up the answers to each item in the scale. We provide details on our procedure in Appendix C, as well as summary statistics and factor loadings on all scale items.

We measure student inputs through three scales of student time spent studying, academic self-efficacy, and mental health, and four additional dummies for whether students display any truant behavior, cheat on exams, aspire to go to university, and expect to be able to go to university. Study effort and initiative are often considered as potential mechanisms for ability peer effects (e.g., Feld and Zölitz, 2017; Xu, Zhang and Zhou, 2020). Our data combines pure study time measures inside and outside the classroom with some indicators of student engagement in their studies. Truancy and exam cheating could be affected positively if higher-ability peers help students perform better in school, relieving some pressure on their studies, or negatively if they become a competitive comparison group causing further stress or marginalization as found in Pop-Eleches and Urquiola (2013). Similar reasons could drive responses in academic self-efficacy and mental health. Non-cognitive skills such as aspirations and expectations about one own ability have also been shown to be partially formed by social interactions and to be productive for academic achievement (Carlana, La Ferrara and Pinotti, Forth.).

We measure parental inputs through four scales capturing private tutoring, time spent with parents, 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. School inputs have been shown to affect parental investments and academic achievement (Pop-Eleches and Urquiola, 2013; Fredriksson, Öckert and Oosterbeek, 2016), and much of this work postulates peer quality as a key driver of these effects (e.g., Pop-Eleches and Urquiola, 2013; Jackson, 2013). Money and time parental investments are the canonical Beckerian household investments in human capital and can therefore respond as complements or substitutes to school inputs, such as peers. Parental strictness, support, and harshness belong to a broader set of parenting styles which have been traditionally modelled in economics as parental investments (Burton, Phipps and Curtis, 2002; Hao, Hotz and Jin, 2008; Cunha, 2015; Doepke and Zilibotti, 2017; Cobb-Clark, Salamanca and Zhu, 2019; Doepke, Sorrenti and Zilibotti, 2019) and thus also react to school inputs for similar reasons. Conceptually, strictness is close to parental monitoring, which can be an important margin in this context, whereas support is closer to warmth and more generally measures parental engagement. Harshness is also not uncommon in this context, and can be a margin of reaction for parents if they see it as a way to ensure their children's school performance. We see parent-child conflict as a potential outcome of all these interactions which can in itself affect student outcomes. Finally, parents' aspirations for their child to go to university can in themselves drive student outcomes (e.g., Janzen et al., 2017) and could proxy for other unmeasured inputs (e.g., parental encouragement).

Lastly, we measure school and teacher inputs through two scales of student-perceived quality of the school environment and of teacher engagement, and two additional dummy variables for whether teachers reports that the classroom is hard to manage and whether they feel tired of teaching. Our index of school environment captures the student's perception of the quality of their school atmosphere, including safety

and study ethos. We could expect students exposed to more competitive classrooms to have a more positive experience in school (Feld and Zölitz, 2017; Booij, Leuven and Oosterbeek, 2017), or to feel marginalized (Pop-Eleches and Urquiola, 2013). Higher-ability peers might decrease classroom disruption, easing classroom management and increasing teacher engagement (Duflo, Dupas and Kremer, 2011). Ultimately, teachers might feel more motivated and less tired of teaching when teaching higher-ability students.

Combined with the mandated random assignment of students to classrooms within schools, we set off to use the richness of these data to explore the effects and mechanisms of having higher-ability peers in junior high school.

3.3 Applying the Fishing Algorithm to the TEPS data

3.3.1 Testing for Random Assignment to Peer Groups

We start by assuming that the Taiwanese quasi-experiment produces the TEPS data ($\mathscr{D}$ in Section 2.1) where the 20,055 students (units n) are assigned to 1,244 classrooms (treatment groups c) across 333 schools (risk sets r) as observed in wave 1. Our focus is on the effects of having higher-ability classroom peers, as measured by average peer test scores, in a linear-in-means model. To that end, we define the treatment as the leave-out mean of classroom peer test scores in wave 1; or, in the terminology of Section 2.1, $T_{ncr} \triangleq \overline{Test\ Score}^{-n}_{ncs1} = (N_{cs}-1)^{-1}\sum_{j\in c, j\neq n} Test\ Score_{jcs1}$ where N_{cs} is the size of classroom c in school s).[8]

To test whether classroom peer ability is truly randomly assigned to students within schools in the data, we consider 18 student characteristics pre-predetermined to peer ability (the set $\mathbf{P}_{ncr}$; see Table 1). Using these, we construct two types of test statistics (the statistic B), both with corresponding critical values from the standard normal distribution in large samples. The first type of test is a t-statistic from a sorting test that relates students' own ability in wave 1 to their classroom peers' mean ability in wave 1 using only within-school variation. We construct two variations of this statistic; the more commonly used one proposed by Guryan, Kroft and Notowidigdo (2009) and its recently corrected version by Jochmans (2023). The second type of test is another t-statistic from balancing tests that relate each of the remaining 17 pre-determined student characteristics to the classroom peers' mean ability, also using within-school variation only.[9]

Panel A of Table 1 shows the Guryan, Kroft and Notowidigdo (2009) and the Jochmans (2023) sorting test statistics, whereas Panel B shows the coefficient and standard error of our balancing tests. Both panels show strong evidence of sorting of students into classrooms within schools when the complete TEPS dataset is considered, in violation of Taiwan's national mandate of random assignments of students

8. Standardized test scores are not strictly measured pre-assignment; they were taken by students within the first month 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 are 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.

9. An alternative balancing test statistic B would be the F-statistic from regressing classroom peers' mean ability on all our pre-determined characteristics and school fixed effects, as suggested by Pei, Pischke and Schwandt (2019). This test leads to the same conclusions.

Table 1: Balancing and Sorting Tests Using the Entire TEPS Data

Panel A: Sorting tests of student test scores on peer test scores			
	Students	*Mean*	*Sorting test t-statistic*
Guryan et al. (2009)	12,793	Std	-0.24
Jochmans (2023)	12,793	Std	-0.69

Panel B: Balancing tests of student pre-assignment characteristics on peer test scores				
	Students	*Mean*	*Peer test scores*	
			Coef.	*Std. err.*
Female student	12,793	0.48	0.005	(0.011)
Student born before 1989	12,741	0.37	-0.014	(0.010)
Household income > NT$100k/mo.	12,600	0.13	-0.020***	(0.007)
College-educated parent(s)	12,278	0.13	-0.003	(0.009)
Parent(s) work in government	12,224	0.09	0.007	(0.007)
Ethnic minority parent(s)	12,276	0.06	-0.004	(0.009)
Prioritized studies since primary school	12,724	0.27	-0.010	(0.009)
Reviews lessons since primary school	12,715	0.18	0.005	(0.009)
Likes new things since primary school	12,688	0.41	0.004	(0.012)
Was truant in primary school	12,632	0.33	0.005	(0.011)
Student had mental health issues in primary school	12,628	0.47	0.004	(0.011)
Had private tutoring before junior high	12,650	0.68	0.010	(0.012)
Family help with homework before junior high	12,166	0.84	-0.023***	(0.008)
Student quarreled with parents in primary school	12,646	0.67	0.004	(0.010)
Student enrolled in gifted academic classroom	12,694	0.06	0.012	(0.008)
Student enrolled in arts gifted classroom	12,694	0.04	-0.011	(0.015)
Parents made efforts to place student in better classroom	12,649	0.15	0.035***	(0.010)

*Estimates in the complete TEPS which includes of up to 333 schools and 1,244 classrooms and 20,055 students. All estimators include school fixed effects. In Panel A, the reference distribution for the Guryan, Kroft and Notowidigdo (2009) and the Jochmans (2023) sorting statistics is the standard normal, and statistics larger than the 10 percent and 5 percent critical values are shown in italics and in bold. In panel B, the rightmost column reports cluster-robust standard errors at the classroom level. ***, ** and * mark estimates statistically different from zero at the 99, 95 and 90 percent confidence level.*

to classrooms within schools.[10]

As discussed in Section 3.1, the law in Taiwan has an explicit and strict mandate of random assignment of students to classrooms, yet the evidence of non-compliance in Table 1 is irrefutable. Through discussions with people familiar with the Taiwanese context, we came to believe that imbalances in classroom assignment within schools are likely driven by punctual non-compliance with the mandate in some schools. Unfortunately, the TEPS does not allow us to infer directly which are these non-compliant schools. This is where our application of the Fishing Algorithm becomes useful.

10. 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 most previous academic peer effect studies. The size of our detected imbalances is 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. Nevertheless, the evidence on non-compliance with random assignment in TEPS is irrefutable.

3.3.2 *Applying the Fishing Algorithm*

When applying the Fishing Algorithm to the TEPS data, we focus on detecting schools that sort students into classrooms based on ability, and we refer to these as 'sorter' schools from now on. Addressing this type of non-compliance is central when estimating ability peer effects. As we show below, once we address ability sorting, the other imbalances become second-order. Below we explain in detail the various decisions we made in applying the Fishing Algorithm in this setting.

In Step 1 of the algorithm, our sorting measure (b_r in Section 2.2) at the risk-set level is the school-level standardized Herfindahl-Hirschman index of test scores which captures the concentration of ability in classrooms for each of the 333 schools in TEPS. We define this as

$$H_s = \frac{HHI_s - \frac{1}{N_s^c}}{1 - \frac{1}{N_s^c}}, \tag{1}$$

where N_s^c is the number of classrooms in school s and HHI_s is defined as

$$HHI_s = \sum_{c \in s} \left(\frac{\sum_{n \in cs} Test\ Score_{ncs1}}{\sum_{n \in s} Test\ Score_{ncs1}} \right)^2. \tag{2}$$

The Herfindahl-Hirschman index is the most prominent measure of market concentration in economics, and our construction measures the concentration of ability (or ability sorting) of student test scores into classrooms within each school, ranging between $1/N_s^C$ if the average test score is identical in all classrooms to 1 if "all high-scoring students are together in one classroom". This may seem extreme with test scores but it can happen for ethnic minority status or gender. The H_s are highly sensitive to variation in concentration, which aids the performance of the algorithm.[11] Intuitively, it is easy to see that the t-statistics of sorting tests presented in Section 3.3.1 are increasing in H_s. This can formally be shown using a variance decomposition to express the numerator of the sorting statistic as the weighted sum of student and peer test score correlation across classrooms, and then taking a partial derivative with respect to this correlation.

In Step 2, we simulate via permutations the random assignment of students to classroom within each school 1,000 times (K in Section 2.2) 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). Each time we construct simulated peer test score measures (the T_{ncr}^k) and use these to construct simulated H_{sk}^{random} (the b_r^k). Each of these reflects one way in which school ability distributions

11. In a different application of the algorithm to the TEPS data, de Gendre et al. (2024) use within-school correlations between the share of Indigenous classroom peers and several pre-assignment non-Indigenous student characteristics as their b_r. Although that choice fits their research question better, the sensitivity of the Fishing Algorithm suffers since these correlations are not penalized non-linearly.

could have looked like if students were truly randomly assigned to classrooms within schools in our observed data. Since we do this 1,000 times, we end up with 333 school-specific ability distributions of H_{sk}^{random} with 1,000 observations approximating each distribution. Finally, we construct for each of the 333 schools the share of instances where the simulated index is larger than the index in the realized data, given by $S_s = 1,000^{-1} \sum_{k=1}^{1,000} \mathbb{1}[H_s > H_{sk}^{random}]$ (the S_r).

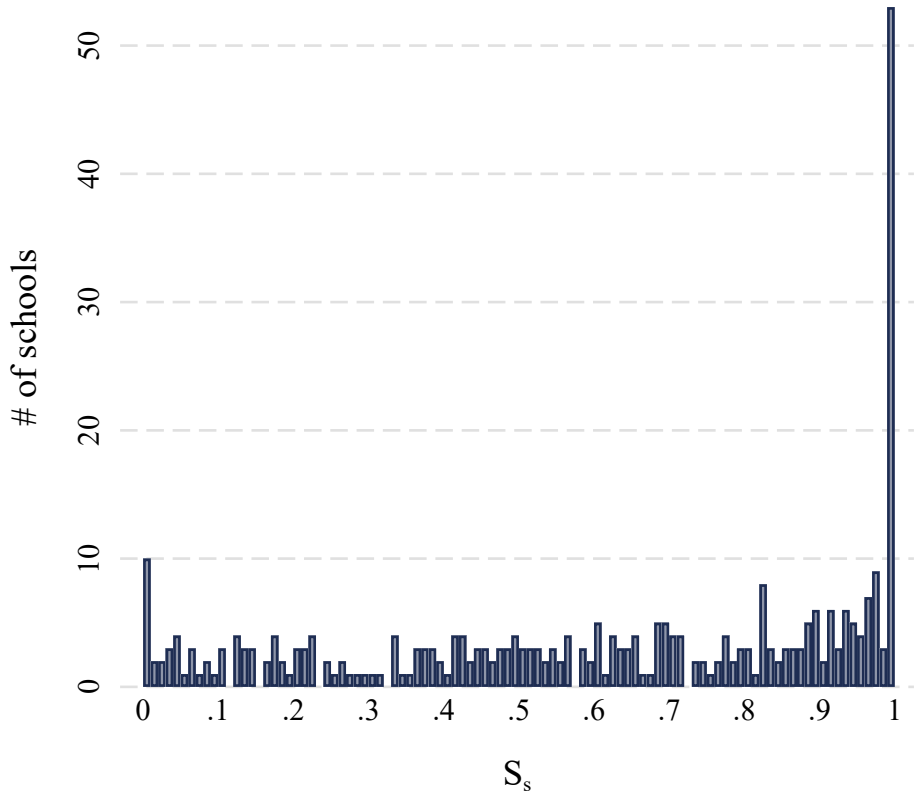
At this point, it is important to highlight why S_s is a better measure of sorting than H_s, especially to measure sorting on characteristics that are relatively rare. Imagine trying to measure sorting based on race in a school with three classrooms and one racial minority student. Even if this school 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 from this example, there are two key lessons. First, S_s naturally normalizes treatment assignment concentrations to reflect risk of treatment at each risk set r, a very useful property. Second, perhaps more subtly, we need a strictly greater number of treated units than treatment groups within a risk set (e.g., more minority students than classrooms in a school) to be able to interpret S_s as a measure of non-compliance with random treatment assignment.

Figure 1 shows the distribution of 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 standard uniform distribution. The figure suggests that most schools are likely complying with the random assignment mandate, yet some schools show a very high degree of ability sorting that is inconsistent with random assignment. After a quick glance at the distribution, one could conclude that schools in the rightmost part of the distribution—with, say, $S_s > 0.9$ which adds up to 47 schools—are much more likely be defying the mandate of random assignment. However, just dropping these 47 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. Bluntly trimming these schools could therefore lead to "over-trimming": removing schools that have high sorting by chance. One problem with over-trimming is that it can lead to negative sorting tests in the trimmed sample. Another problem is that it would remove legitimate variation from the estimation sample that could be crucial for identifying peer effect. In the worst-case scenario, this could result in a loss of power for identifying peer effects in the trimmed sample and, if peer effects are extremely non-linear, over-trimming could bias peer effect estimates downwards. This power issue is especially concerning for *a priori* small effects such as peer effects. Hence the importance of Step 3 in our Fishing Algorithm.

In Step 3, there are three key choices to make: i) the correct model given the distribution of S_s (the function $f_i(\cdot)$ in Section 2.2), ii) the number of latent classes (the L), and iii) the school-level class predictors, if any (the $\mathbf{Z_r}$ regressors).

Modelling the distribution of S_s: We opted for fitting a right-censored Finite Mixture Model (FMM) Tobit to account for the 34 schools that show censoring of S_s at 1. We attempted to model the point-mass at 1 as a separate class, akin to how FMMs are used to fit zero-inflated negative binomials, but these models had convergence issues. Explicitly accounting for the censoring of seven schools at zero (either

Figure 1: The Distribution of S_s Measuring Sorting of Students into Classrooms Within Schools Based on Their Test Scores in Wave 1



This figure shows the school-level distribution of S_s which measures whether schools sort students into classrooms based on their test scores in Wave 1 of the TEPS more strongly than chance would allow, given the school size, number and classroom size and student composition. This distribution was calculated by comparing actual and simulated standardized Herfindahl-Hirschman indices H_s for each school using 1,000 simulated random assignments of students to classrooms within schools.

by using a left-censored Tobit or by modelling the point-mass separately) is empirically inconsequential, yet it makes convergence harder.

Choosing the number of latent classes: We chose the one that minimized the Akaike and Bayesian Information Criteria, which in our case led to a 3-class model. In Table 2 we show key results of this 3-class model, with Panel A showing the estimated marginal mean of S_s ($\hat{\mu}_i$ in Section 2.2), which is central for labeling the classes. Panel A also shows the variance of S_s for each of the three latent classes as well as the estimated class posterior probabilities. The first latent class describes schools that comply with the random assignment mandate (with a marginal mean of S_s close to 0.5, a large variance, and capturing almost 60 percent of schools), the second class describes sorter schools (with a marginal mean close to 1, a very small variance, and capturing 28 percent of schools), and the third class describes "balancer" schools—in which student ability is more evenly distributed across classrooms than predicted by chance—effectively capturing the small probability mass at zero in Figure 1 (with a marginal mean close to 0, a small variance, and capturing the remaining 15 percent of schools). Models with two and four latent classes lead to very similar results, with one clearly identifiable sorter latent class that classified nearly all the same schools as sorters. The fact that the balancer latent class is also identified in our data

is consistent with a few schools implementing the S-sequence grouping described in Section 3.1, though this grouping seems somewhat less common.

Choosing school-level class predictors: We found that models with many of these covariates almost never converged, so we reduced our set of covariates to those most predictive of S_s via least absolute shrinkage and selection operator (LASSO) regression of S_s on a large set of potential school-level covariates, with the regularization parameter value λ^* chosen via 10-fold cross-validation. This procedure selected the following seven covariates out of a potential set of 66: schools means for ethnic minority students, for whether students review lessons with their parents in primary school, whether students report being in academically gifted classrooms for arts or for other subjects, and three school area dummies for Taipei City, Hsinchu County, and Yunlin County. A likelihood ratio test confirms that adding these school-level class predictors significantly improves the model fit (with p-value < 0.001 in every instance).

Table 2: Estimates of a Three Latent Class Finite Mixture Model of S_s

Panel A: Post-Estimation Statistics of a 3 Latent Class Model of S_S			
Model latent class =	*1*	*2*	*3*
Latent class marginal mean	0.598	0.979	0.097
	(0.028)	(0.009)	(0.015)
Latent class variance	0.051	0.003	0.006
	(0.009)	(0.001)	(0.002)
Posterior latent class probabilities	0.567	0.282	0.151
	(0.034)	(0.027)	(0.019)
Panel B: Class Predictor Coefficients and Standard Errors (Base: Class 1)			
Ethnic minority parent(s)		-0.205	2.490**
		(1.543)	(1.243)
Reviews lessons since primary school		-1.684	-11.163**
		(2.819)	(4.337)
Student enrolled in gifted academic class		6.852***	-1.873
		(2.090)	(3.632)
Student enrolled in arts gifted class		9.934***	4.214
		(2.581)	(2.979)
Taiwan administrative area (base: all others)			
Taipei City		0.607	-0.088
		(0.433)	(0.553)
Hsinchu County		-39.527	0.364
		(2.660e+08)	(0.997)
Yunlin County		1.837	-17.983
		(1.198)	(14912.865)
LR test for class predictors [p-value]		<0.001	
Schools		333	

*Estimates from a Finite Mixture Model (FMM) with 3 latent classes estimates in the complete TEPS sample of 333 schools using school-level data. The dependent variable is S_s, modeled via a Tobit with right censoring at 1. A parsimonious set of regressors are selected via a least absolute shrinkage and selection operator (LASSO) of the dependent variable on all school-level means of pre-assignment variables and a complete set of dummies for school geographical characteristics (reducing the number of regressors from 66 to 7). 100 random draws of posterior probabilities are used to search for initial values to improve convergence. The 3-class model is preferred to a 2-class model by both AIC and BIC criteria. ***, ** and * mark estimates statistically different from zero at the 99, 95 and 90 percent confidence level.*

A key benefit from Step 3 in our Fishing Algorithm is revealing the school-level characteristics that predict whether a school is a sorter. This characterization comes directly from the FMM estimates; in particular,

from the coefficient estimates on predictors of school s belonging to the latent class of sorter schools (the $\hat{\psi}_l$ in Section 2.2). Panel B of Table 2 shows these estimates, with the class of compliant schools as base. We focus on the middle column, which shows that the strongest predictors of belonging to the sorter school class are the school shares of students that report being in an academically or artistically gifted class. Their exponentiated coefficients correspond to relative risk ratios. They imply that a 10 percentage points increase in the share of students reporting being in an academically (arts) gifted classroom increases the chance that a school is classified as a sorter rather than a complier by almost twice (2.7 times). While these magnitudes seem large, note that on average only 7.9 percent (5.6 percent) of students report being in an academically (arts) gifted classroom. These associations strike a close parallel with circumstantial evidence we later uncovered suggesting that schools use off-the-books gifted classrooms as a tool to bypass grouping regulations.[12] The formation of gifted classrooms is also heavily regulated and can in principle only be done with special authorization from the government, yet it seems some schools do it regardless. In fact, two data features revealed by our method are consistent with the off-the-books gifted classroom hypothesis: (1) that the share of students in a school reporting being in a gifted class is a strong predictor of sorter schools, and (2) that the principal-reported presence of gifted classrooms in the school (which was also among the potential set of covariates) *is not* a good predictor of sorter schools.[13]

In Step 4, we flag schools as sorters using the rule $\tilde{\pi}_{lr} > \tau = 0.5$ as suggested in Section 2.2. We therefore classify as sorter schools those which were more likely to belong to the sorter latent class than to all other classes combined, according to their predicted posterior probabilities. Different thresholds τ can be justified, but this is a reasonable one with clear *a priori* basis. With this rule we identify 102 out of the 333 schools in TEPS as sorters. In itself, this result provides evidence that our suspicions on imbalances being driven by only some sorter schools were indeed justified. If imbalances were driven by generalized sorting behavior across most schools, our algorithm would have produced very different results (see Case 3 in Appendix A).

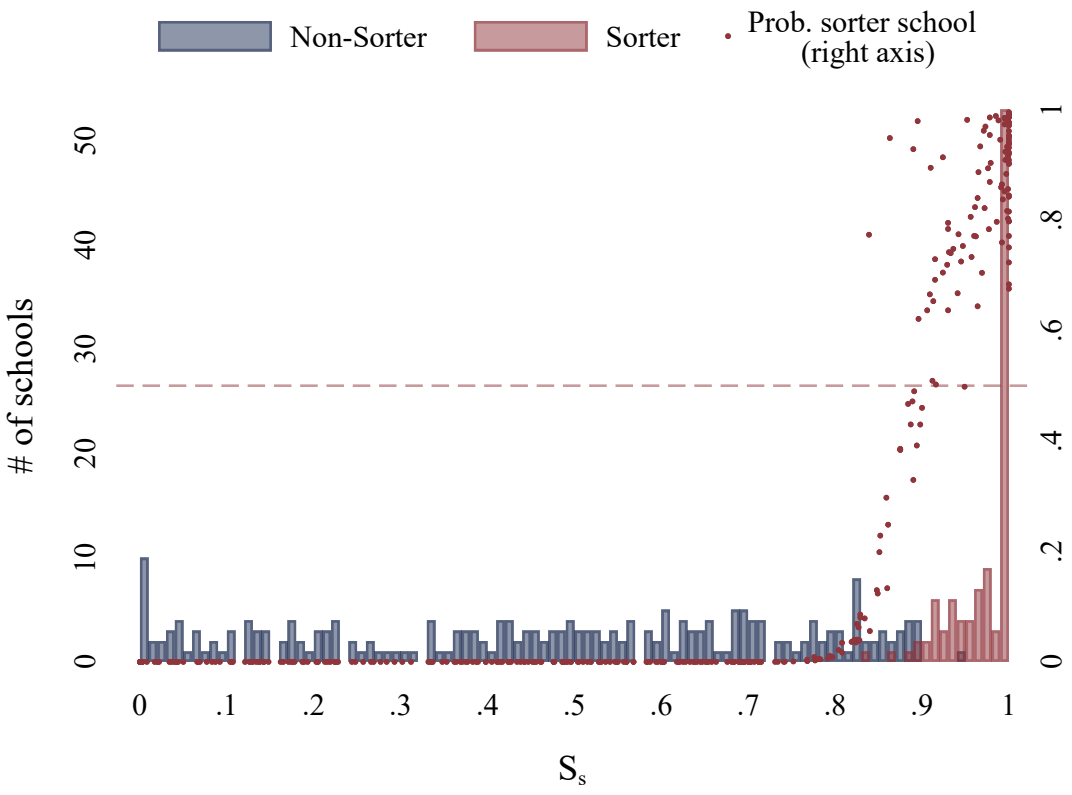
Figure 2 shows the schools eventually flagged as sorters by our Fishing Algorithm across the distribution of S_s in maroon. We overlay $\tilde{\pi}_{lr}$, the predicted posterior probability of being a sorter school, on the right y-axis in a scatterplot, with 0.5 as the dashed horizontal reference line. Schools flagged as sorters are also marked in maroon bars in the distribution of S_s. As expected, most flagged schools have $S_s > 0.9$, though a few schools with lower values of S_s are also flagged. In the TEPS data, the algorithm classified all schools with high S_s as sorters. It is possible, of course, that all these schools are indeed sorters, yet it is more likely that the FMM class predictors are just not strong enough to discern compliant schools among this group that simply happen to have a high S_s by chance. As discussed above, this could lead to

12. See the article "Mandatory eighth period classes, morning and evening self-study, and disguised ability-based classes... How far are we from true educational equality?" published in *The News Lens* on January 2, 2022 and available online at https://www.thenewslens.com/article/161761 (in Mandarin, last consulted on 13 March 2024). This article documents the existence of *de facto* gifted classrooms that combine the top-performing students in each classroom. These classrooms often get additional resources, such as better teachers or more teaching time. The article also speculates on the existence of behind-the-scene adjustments to lot-drawing or random number lotteries. We cannot observe the latter using our data.

13. The rightmost column in Panel B of Table 2 further suggest that schools with more minority students are more likely to be compliant schools. The former effect could be because in Taiwan schools with a meaningful share of Indigenous students are classified as Indigenous schools under the *Education Act for Indigenous Peoples*, and therefore are subject to additional regulation and likely further scrutiny. Under those conditions, S-sequence grouping could be more defensible since it rules out chance imbalances across ethnicity. We have no working model for the other coefficients in Panel B.

over-trimming and in fact we do see some of it, which we discuss below. Yet evidence of over-trimming is not strong enough to be concerning.

Figure 2: Schools Identified as Systematically Sorting Students into Classrooms by Student Academic Ability Using the Fishing Algorithm



This figure shows the school-level distribution of S_s which measures whether schools sort students into classrooms based on test cores more strongly than chance would allow, given the school size, number and classroom size and student composition. The probability of being a sorter school (i.e., a school that sorts students into classrooms based on academic ability), $\hat{P}_{sl}$, is estimated as the posterior probability of being in a latent class classified as sorters *by us based on a Finite Mixture Model of S_s with three latent classes and using several school averages of parental characteristics as class predictors (see Appendix Table 2). $\hat{P}_{sl}$ are shown as maroon dots on the figure and scaled on the left y-axis. For reference, the 0.5 probability is shown as a dashed horizontal line. Schools where $\hat{P}_{sl}$ exceeds 0.5 are classified as sorters and displayed in maroon in the figure. The rest of the schools are displayed in navy blue.*

We then exclude the 102 schools classified as sorters from our data, leaving us with a trimmed subsample of 13,750 students in the remaining 231 schools, allocated to 851 classrooms ($\mathscr{T}$ in Section 2.2, which in this application covers 69 percent of the TEPS data) where the Taiwanese quasi-experiment should hold. Table 3 presents the results of sorting and balancing tests on this trimmed sample, following the same format as Table 1. Panel A shows that both the Guryan, Kroft and Notowidigdo (2009) and Jochmans (2023) fail to detect any ability sorting in this sample. Panel B shows little evidence that unbalancedness is a concern in our trimmed sample. Of all 17 pre-assignment characteristics we test, only three characteristics are statistically significantly related to peer test scores in wave 1. Household income and family engagement with homework before baseline are *negatively* related to peer test scores. This last finding rather suggests a potential slight over-trimming in our Fishing Algorithm (since in the complete TEPS data these relationships were, if anything, positive; see Appendix Table 1). There is a

positive relationship between whether parents made efforts to have their child assigned to a better class and peer test scores, which suggests that pushy parents might have had a measure of success, even in this trimmed sample. However, the magnitude of the coefficient is small—a 10 percentage point increase in pushy parents in the classroom is associated with a mere 0.35 percent of a standard deviation in baseline test scores. In any case, in ourof our preferred specifications below we include controls for these three characteristics as well as a host of additional controls. These are not crucial for our empirical design and their inclusion never affects any of our main results.[14]

Table 3: Sorting and Balancing Tests of Peer Test Scores on the TEPS Trimmed Sample

Panel A: Sorting tests of student test scores on peer test scores				
	Students	*Mean*	*Sorting test t-statistic*	
Guryan et al. (2009)	12,793	Std	-0.24	
Jochmans (2023)	12,793	Std	-0.69	
Panel B: Balancing tests of student pre-assignment characteristics on peer test scores				
	Students	*Mean*	*Peer test scores*	
			Coef.	*Std. err.*
Female student	12,793	0.48	0.005	(0.011)
Student born before 1989	12,741	0.37	-0.014	(0.010)
Household income > NT$100k/mo.	12,600	0.13	-0.020***	(0.007)
College-educated parent(s)	12,278	0.13	-0.003	(0.009)
Parent(s) work in government	12,224	0.09	0.007	(0.007)
Ethnic minority parent(s)	12,276	0.06	-0.004	(0.009)
Prioritized studies since primary school	12,724	0.27	-0.010	(0.009)
Reviews lessons since primary school	12,715	0.18	0.005	(0.009)
Likes new things since primary school	12,688	0.41	0.004	(0.012)
Was truant in primary school	12,632	0.33	0.005	(0.011)
Student had mental health issues in primary school	12,628	0.47	0.004	(0.011)
Had private tutoring before junior high	12,650	0.68	0.010	(0.012)
Family help with homework before junior high	12,166	0.84	-0.023***	(0.008)
Student quarreled with parents in primary school	12,646	0.67	0.004	(0.010)
Student enrolled in gifted academic classroom	12,694	0.06	0.012	(0.008)
Student enrolled in arts gifted classroom	12,694	0.04	-0.011	(0.015)
Parents made efforts to place student in better classroom	12,649	0.15	0.035***	(0.010)

*This table shows results of sorting and balancing tests on peer test scores in our trimmed sample of up to 231 schools and 851 classrooms and 12,793 students. All estimators include school fixed effects. The reference distribution for the Guryan, Kroft and Notowidigdo (2009) and the Jochmans (2023) sorting statistics is the standard normal. The last column of Panel B reports cluster-robust standard errors at the classroom level. ***, ** and * mark estimates statistically different from zero at the 99, 95 and 90 percent confidence levels.*

Overall, our Fishing Algorithm is an effective way to identify schools that systematically assign student

14. Note that due to the power in our data, we can detect small differences in balancing tests that would have likely gone unnoticed in other datasets. In our balancing tests, for example, our ex-post Minimum Detectable Effects (MDEs) assuming 80 percent power and 95 percent confidence are as small as 3.1 percentage points in the chance of being female and 2.5 percentage points in the likelihood of having a migrant background. In comparison, in the Add Health data the MDE of balancing tests for being female are 25 percentage points and are 10 percentage points for migrant (Bifulco et al., 2014).

to classrooms in our data, and to recover the subset of schools compliant with the random assignment mandate. In the schools identified as compliant, we find no substantive evidence of systematic assignment. We keep this trimmed sample as our estimation sample throughout the remainder of our analyses. In Section 3.5 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.[15]

Now that we have established a sample where conditional random assignment of students to classrooms within schools holds in our trimmed data, we use these data to explore the existence and mechanisms of ability peer effects in Taiwan.

3.4 Ability Peer Effects on Test Scores and New Evidence on Mechanisms

3.4.1 Ability Peer Effects on Test Scores

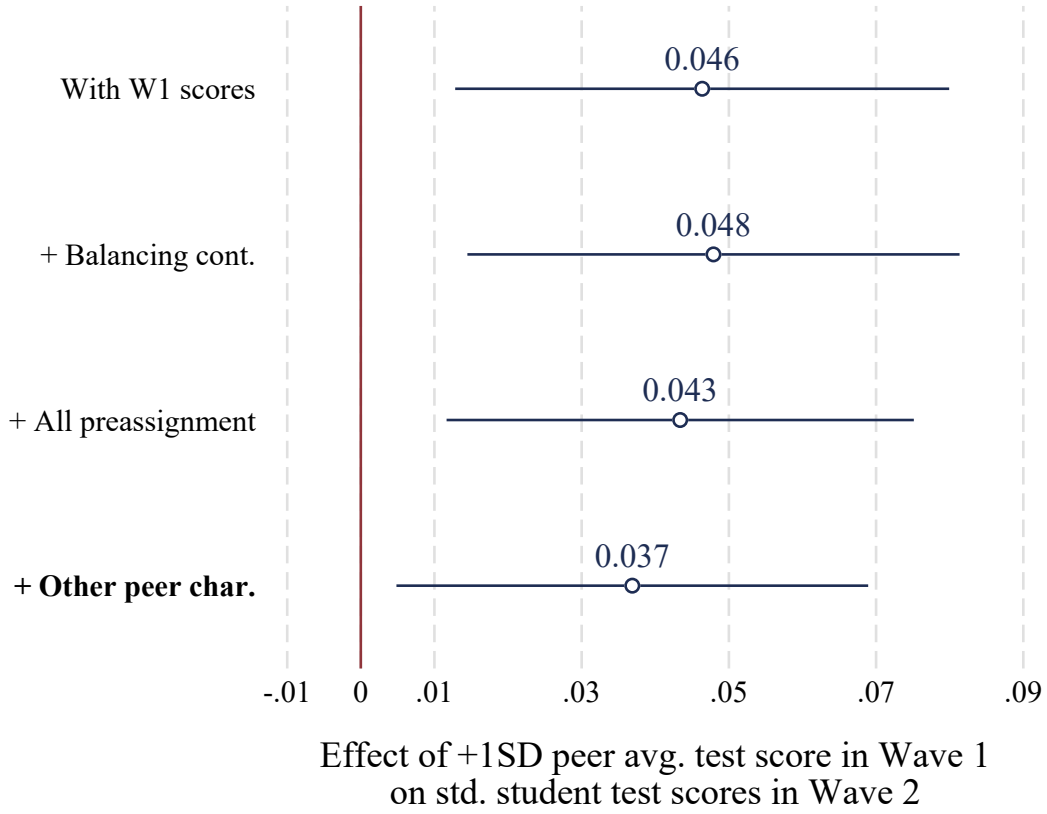
In its most basic form, we estimate ability peer effects in our setting by regressing students' standardized test scores in wave 2, $Test\ Scores_{ncs2}$, on the standardized classroom leave-out mean of test scores in wave 1, $\overline{Test\ Scores}^{-n}_{ncs1}$, our measure of average peer test scores. To this simplest specification we add school fixed effects and students' own test scores in wave 1. We consider specifications with balancing controls (household income, family engagement with homework and parents' pushiness to get child assigned to a particular classroom), with controls for all remaining 14 pre-assigned characteristics (student gender and year of birth, educational attainment and occupation of parents, parental ethnic background, whether the student reports being in an academic or arts gifted classroom, behavior and health during primary school, attitudes towards school since primary school, and family investments in tutoring and homework before junior high school), and finally with controls for peer averages of all these pre-assignment characteristics. This last set of control variables ensures that the ability peer effects we identify can truly be ascribed to higher peer ability rather than to other peer characteristics correlated with ability. We do not include homeroom teacher fixed effects since these teachers are also randomly assigned to classrooms so they cannot confound our peer effect estimates (see Section 3.1 and Chang, Cobb-Clark and Salamanca, 2022). Moreover we do not observe the same homeroom teacher across multiple classrooms so our estimates would not be econometrically identified in a teacher fixed effect model. We cluster standard errors at the classroom level.

The results in Figure 3 (also reported in Appendix Table B.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 and after trimming 102 schools.

Our preferred specification is on the last row of Figure 3, highlighted in bold, and can be re-expressed as:

15. As a side note, we show that out applying our Fishing Algorithm in the TEPS data does not introduce any evident selectivity in our estimation samples. Table B.1 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 sorters by our Fishing Algorithm.

Figure 3: The Effect of Peer Test Scores in Wave 1 on Students' Own Test Scores in Wave 2



This figure reports coefficient estimates of regressing standardized student test scores in wave 2 on standardized average peer test scores in wave 1 in our trimmed sample of up to 231 schools and 851 classrooms and 12,793 students. Estimates in this figure are also shown in Appendix Table B.3. Rows present results of models with increasing sets of control variables going down. W1 scores stand for student's own test scores in wave 1. Balancing cont. stands for pre-assignment control characteristics unbalanced in wave 1 (household income, family help with homework and whether parents tried to influence junior high classroom assignment). All pre-assignment stands for all 17 pre-assignment characteristics tested for balancing. Other peer char. stands for the leave-out mean of other peer characteristics including gender, age, family income and education, government-employed parents, ethnic minority status, and reports of attending gifted classes, or having pushy parents. All models include school fixed effects. Missing covariates are imputed at the median and a missing covariate flag is always added. Horizontal bars show the 95 percent confidence intervals for each estimate, based on standard errors clustered at the classroom level.

$$TestScores_{ncs2} = \underset{(0.016)}{0.037}\,\overline{TestScores}^{-n}_{ncs1} + \underset{(0.007)}{0.710}\,TestScores_{ncs1} + \hat{\Gamma}'Controls_{ncs1} + \hat{\gamma}_s \tag{3}$$

where $Controls_{ncs1}$ includes student and peer pre-assignment characteristics, and $\hat{\gamma}_s$ represent school fixed effects that are partialled out of the variation identifying all coefficient estimates.[16]

These estimates imply that having one standard deviation higher average peer test scores in wave 1 increase own test scores by 3.7 percent of a standard deviation in wave 2. Comparing effect sizes in

16. This is, of course, not the only way we could have modelled peer effects—though our linear-in-means specification is certainly among the most popular. Ultimately, we focus on the linear-in-means model due to its strong behavioral micro-foundations (Ushchev and Zenou, 2020; Boucher et al., 2024), because it is the workhorse of empirical peer effect models in the literature, and because it is parsimonious yet captures peer dynamics very well in our data.

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 the variance of peer ability measures 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; Booij, Leuven and Oosterbeek, 2017). Compared to studies where students are randomly assigned to peer groups, our estimates are around the median of estimates. 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.[17]

To give this number more perspective, our estimated effect of a 1SD increase in average peer test scores is around five percent of the estimated impact of a 1SD increase in students' own lagged test scores. Our peer effect estimate is about a fifth of the marginal effect of having at least one college-educated parent, and about a tenth of the unconditional test score gap between children of two-parent households and single-parent households.[18]

Having established the effect of higher-ability peers on test scores in our data, we now proceed to estimate their impact on 18 educational inputs which comprise key underlying mechanisms behind ability peer effects.

3.4.2 Ability Peer Effects on Educational Inputs

We fit variations of Equation (3) using our measures of educational inputs in wave 2 as outcomes. Figure 4 (also reported in Appendix Table B.4) shows the effect of a 1SD increase in average peer test scores on wave 2 educational inputs using our preferred specification. 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 green estimates show effects on school and teacher inputs. Horizontal spikes around point estimates mark 95 percent confidence intervals.

We manage to rule out effects through 14 out of 18 inputs in our data via precisely estimated null effects. These include null effects on all seven student inputs, in five out of seven parental inputs, and in two out of four school and teacher inputs we measure. Far from being uninformative, however, these null effects provide valuable information to a literature where there is much speculation but scant evidence on mechanisms. The combination of many potential mechanisms and little evidence also means that the literature provides weak support for a large prior belief on the effect of peer ability on any one mechanism for any one point null. Abadie (2020) shows that it is precisely in these situations where null effects can

17. The combination of partial classroom sampling and random assignment of students to classes in TEPS implies that these and all other peer effect estimates in our main results might be biased towards zero (Sojourner, 2013). We discuss the source of this bias, and present and interpret corrected estimates, in Section D.2.2.

18. Another way of sizing the impact of higher-ability 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 poorest 10 percent in the sample) have 1.04 standard deviations lower scores than the peers of rich students (with household monthly incomes over NT$100,000, corresponding to the top 15 percent). The rich-poor test score gap in wave 2 test scores gap is around 99 percent of a standard deviation. Putting these two numbers together, our linear peer effects imply that 3.6 percent of the rich-poor gap in standardized test scores can be explained by the richer students' access to peers with higher test scores.

result in large shifts in prior beliefs, and therefore be most informative.

Study effort, for example, is probably the most-often considered mechanisms in virtually every peer effect paper, yet only a handful of them provide estimates of effort responses to peer test scores. Consistent with our findings, the few studies that do estimate effect on effort find mostly nothing. Feld and Zölitz (2017) find no impact of tutorial peer GPA on self-reported study hours, and Fang and Wan (2020) argue that peers' study effort does not explain the effect of exposure to higher-ability roommates on test scores. Although they focus on exposure to female peers, Lavy and Schlosser (2011) also find no impact on study effort.[19] Our null effects on academic self-efficacy also contrast with Zárate (Forth.), who reports negative effects of higher-achieving peers on students' perception about their own abilities and on their self-confidence. However, in contrast with our findings and with most of the literature, Zárate (Forth.) also reports *null to negative* effects of higher-ability peers on academic achievement of girls.

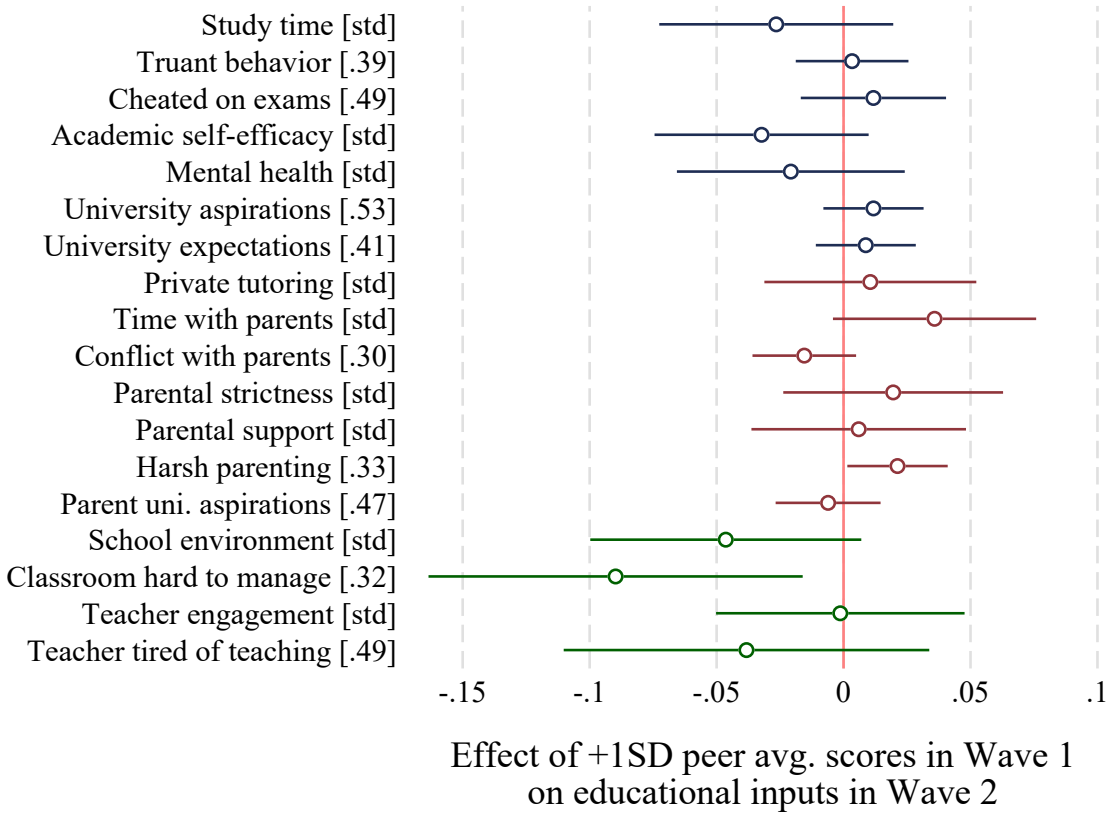
Our null effects on all these potential mechanisms are precisely estimated, which further adds to their informativeness (see Abadie, 2020, p. 200). Between all our estimates, the largest standard error for a standardized educational input is relatively small, at 0.027. A standard ex-post Minimum Detectable Effect (MDEs) size calculation with 95 percent confidence and 80 percent power implies that we could have detected effects as small as 7 percent of a standard deviation for teacher engagement and as small as 6 percent of a standard deviation for private tutoring. These are small detectable effects; they are smaller than 10 percent of the gender gap in effort (girls pay more effort than boys), 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.

We do nonetheless find evidence supporting two mechanisms and suggesting another two mechanisms. First, we find that a 1SD increase in peer test scores results in a large reduction of teachers' reports that the classroom is hard to manage by 9 percentage points, which is 28 percent of the unconditional mean. Second, we also find a small increase in the use of harsh parenting by 2.1 percentage points, which is 6.5 percent of the unconditional mean. The former effect has been considered and, to some extent, tested in the form of classroom disruption (e.g., Lavy and Schlosser, 2011; Feld and Zölitz, 2017; Lazear, 2001; Duflo, Dupas and Kremer, 2011; Golsteyn, Non and Zölitz, Forth.). The latter finding is new and may seem somewhat surprising. However, there is good evidence that parents often consider classroom peers to be an important measure of the quality of their child's school environment (e.g., Abdulkadiroğlu et al., 2020; Beuermann et al., 2023). If Taiwanese parents also treat classroom peers as an input educational input and they see harsh parenting as a form of parental investment (see e.g., Doepke and Zilibotti, 2017; Cobb-Clark, Salamanca and Zhu, 2019), our findings would indicate that Taiwanese parents complement school inputs with their own investments. This new finding contributes to the literature on parental responses to educational inputs, and is consistent with previous evidence from Taiwan (Chang, Cobb-Clark and Salamanca, 2022).

We also find some marginally significant evidence that higher peer test scores increase time spent with

19. Somewhat in contrast, Mehta, Stinebrickner and Stinebrickner (2019) find that students exposed to roommates with a higher propensity to study – and not to higher-ability peers – increase their own study effort, which ultimately results in higher test scores.

Figure 4: The Effect of Peer Test Scores in Wave 1 on Educational Inputs in Wave 2



This figure reports coefficient estimates of standardized average peer test scores in wave 1 on measures of student, parent, and teacher educational inputs between waves 1 and 2 (measured in wave 2) in our trimmed sample of up to 231 schools and 851 classrooms and 12,793 students. Estimates in this figure are also shown in Appendix Table B.4). Different rows correspond to different educational inputs. Squared brackets mark inputs measured as standardized indices (with a mean of zero and a standard deviation of one) as 'std', or report the unconditional mean of binary inputs. All models control for student's own test scores in wave 1, all 17 pre-assignment characteristics tested for balancing, and the leave-out mean of peer characteristics. Missing covariates are imputed at the median and a missing covariate flag is always added. Horizontal spikes show the 95 percent confidence intervals for each estimate, based on standard errors clustered at the classroom level.

parents by 3.6 percent of a standard deviation and decrease student's ratings of the quality of the school environment by 4.6 percent of a standard deviation. The time investment measure in TEPS focuses on dinner time spent with parents and our estimated positive effect is small. It compares to a fifth of the impact of having one student more in one's classroom on parents' likelihood of helping the child with homework in Fredriksson, Öckert and Oosterbeek (2016), or with less than a tenth of the effect of a child attending a marginally worse school in Pop-Eleches and Urquiola (2013). The negative effect on school environment is consistent with the idea that classrooms with higher-ability peers are more competitive, making the study atmosphere tense and unpleasant. Our result is consistent with Pop-Eleches and Urquiola (2013), who find that pupils marginally admitted to selective schools are more likely to feel socially isolated and to be victimized, and with Booij, Leuven and Oosterbeek (2017), who find that lower-ability undergraduate students tracked with similar ability students enjoy more positive interactions with their classmates, compared to lower-ability students in mixed-ability groups. However, our result contrasts with Feld and Zölitz (2017), who find that undergraduate students exposed to higher-achieving tutorial peers report better group functioning. In any case, we are cautious about our interpretations of

these effects since they are less precisely estimated and can reflect substantial sampling error.

Overall, we show that higher-ability peers reduce teacher's reports that the classroom is hard to manage, increase parents' use of harsh parenting, and may also increase parents' time investment and decrease students' ratings of the quality of the school environment. Higher-ability peers also have precisely estimated null effects on every other mechanism we considered. Together, the results of this application advance the knowledge frontier on the mechanisms behind ability peer effects on several margins, while also illustrating the difficulties of learning about mechanisms. Our findings also 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 summarize evidence that our main results and conclusions are robust to a myriad of modeling choices and potential concerns.

3.5 Sensitivity Analyses and Additional Results

Appendix D presents in detail the sensitivity of our results along three dimensions. First, we perform sensitivity analyses related to the Fishing Algorithm and our identification strategy. We show balancing using less common but more comprehensive balancing tests. We then show that our main results hold under more and less stringent parameters for trimming sorter schools. Last, we also show that using estimated sorter probabilities as weight can further bring efficiency gains without introducing measurable bias.

Second, we explore the sensitivity of our results to measurement error in our data. We show that our main estimates are robust to using different measures of student and peer academic ability, that our estimates are are not attenuated by measurement error in average peer test scores, and are, if anything, biased towards zero by the fact that we do not observe entire classrooms.

Third, we estimate the robustness of our inference to different ways to account for sampling error, and show that inference on our results is robust to constructing standard errors based on recent randomization inference techniques, but that the effects on educational inputs do not survive the corrections for multiple hypotheses testing.

Finally, we present four additional sets of results. We show there is little heterogeneity in our ability peer effects on test scores across several dimensions. Using ancillary survey data, we also show suggestive evidence of the effect of higher-ability peers on long-term outcomes. We then combine our estimates of higher-ability peer effects on test scores and educational inputs with value-added models to show that our results imply a null mediated effect. This is not surprising given that we find scant evidence on mechanisms. However, we end by showing that higher-ability peers cause the changes in the value-added of different educational inputs, which suggests peer-induced technology changes as a new type of potential mechanisms for our ability peer effects.

4 Conclusions

We develop the Fishing Algorithm, a data-driven method to detect local non-compliance with a random treatment assignment rule in specific risk sets, and we demonstrate how to use our method to recover valid quasi-experiments from existing data. Our method is easily generalizable across contexts, requires little institutional knowledge other than that to identify well-defined risk sets in the data, and can jointly classify and characterize non-compliant risk sets. We then illustrate the usefulness of our method by estimating ability peer effects and their mechanisms in Taiwan, where we have unusually rich data and where there is random assignment of students to classrooms within school, though this assignment rule is not always respected. After using the Fishing Algorithm to detect and remove from our data the set of schools that do not comply with the random assignment mandate, we use the remaining data to estimate ability peer effects and 18 potential mechanisms. We find effects commensurate with other studies where peers are randomly assigned, and we find evidence supporting four out of 18 potential mechanisms in what is the most comprehensive investigation of peer effect mechanisms to date.

One potential limitation of our method is that it works best when there is good information at the risk set level. Specifically, the method excels when the researcher has access to good predictors of whether a risk set is likely non-compliant (i.e., variables at the risk set level that are highly correlated to S_r). This will be the case in many settings, since the implementation of experiments and the exploitation of quasi-experiments often requires researchers to collect a lot of information about and around risk sets. In our application risk sets are schools and treatment groups are classrooms. But when such information is not available or is of low quality, the Fishing Algorithm would not allow researchers to discern the observable characteristics associated with treatment assignment non-compliance.

Methodologically, the most promising avenues for improvement relate to our reliance on finite mixture models. A first way to enhance the Fishing Algorithm could be to fine-tune the finite mixture models we use to better respect the limited nature of their dependent variable, i.e., its $[0, 1]$ support with likely point masses at both ends of this range. To this end, adapting likelihood functions from fractional regression models seems like a natural candidate. This might lead to efficiency gains, though in a brief exploration of this improvement we found that it worsened convergence issues. One could also use methods other than finite mixture models in order to classify and categorize non-compliant risk sets. Potential alternatives include hierarchical clustering, support vector machines (SVM), and deep neural networks. All these alternatives present benefits and drawbacks.

Regarding our application, we highlight that the Fishing Algorithm permitted the most comprehensive study of the effect of exposure to high-achieving peer on educational investments. We are the first to explore so many potential mechanisms of achievement peer effects, a topic where researchers have found notoriously little evidence in over twenty years of research and hundreds of articles. Our results 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 classroom disruption, increasing stress and deteriorating mental health for both students and teachers, and higher effort to keep up with one's higher-achieving peers. This is a central problem for the design of effective classroom assignment policies (e.g., Fruehwirth, 2014). Our study suggests that these concerns

are unfounded, at least this context. If anything, higher-achieving peers seem to make it easier for teachers to manage classrooms, an effect that echoes previous findings in e.g., Feld and Zölitz (2017). In the absence of measurable costs, our results suggest that higher-achieving peers could be an effective way to get small increases in student achievement, even if we do not fully understand how they work yet. To bring in such changes via classroom reassignment policies would, of course, require a careful exploration of non-linearities in achievement peer effects (Graham, 2008). Yet our results suggest that this type of externalities from higher-achieving students can help explain why better peers make for better schools, as discussed in Cullen, Jacob and Levitt (2006); Pop-Eleches and Urquiola (2013); Jackson (2013); Fredriksson, Öckert and Oosterbeek (2016) and Bütikofer et al. (2020), among others.[20]

We see at least two avenues in which future research on ability peer effects can build on our results. The first one is to keep on striving to find data on potential mechanisms. It is true that most potential mechanisms for achievement peer effects proposed in previous studies feature in one way or another in the TEPS, many of them more carefully measured than ever before. Three notable exceptions are direct learning from peers, detailed teaching practices and endogenous friendship formation. Measuring direct peer learning (e.g., discussing tasks and coordinate among group mates) requires data on peer-to-peer interactions which is difficult to gather, yet could indeed be the missing explanation for achievement peer effects (e.g., Garlick, 2018; Zárate, Forth.; Kimbrough, McGee and Shigeoka, Forth.). Detailed teaching practices (e.g., how teachers pair students for group work or the amount of material covered in each lesson) are also hard to measure yet some of them are strongly related to student achievement gains (e.g., Kane et al., 2011) and one can easily think of ways in which teachers adapt their teaching style to classroom ability. Endogenous friendship formation is the key mechanisms suggested by Carrell, Sacerdote and West (2013) and we are still missing hard evidence supporting this hypothesis. The second avenue is to explore whether, instead of changing inputs, higher-achieving peers change the technology of learning itself. That is, whether students who share a class with higher-achieving peers experience more learning for the same amount of e.g., study time, class engagement, or parental investments. We find some evidence that this could be the case (see Appendix Table B.16). Detailed data on peer-to-peer interactions will also be crucial for further exploring this possibility.

20. One might wonder if our results can really tell us something about achievement peer effects in other settings. Data from the Trends in International Mathematics and Science Study (TIMSS) in 1999 shows that Taiwan's educational setting is altogether not that different from many others across the world (Appendix Table B.17). Taiwan is comparable to other countries—especially in the Australasia and Pacific region—in most other key dimensions including class size, student-teacher ratios, daily study hours, dropout rates, or class disruption. Although students in Taiwan do spend relatively more days per year in school, and have lower rates of absenteeism, we find positive achievement peer effects of similar size to many other studies.

References

Abadie, Alberto. 2020. "Statistical nonsignificance in empirical economics." *American Economic Review: Insights*, 2(2): 193–208.

Abdulkadiroğlu, Atila, Parag A Pathak, Jonathan Schellenberg, and Christopher R Walters. 2020. "Do parents value school effectiveness?" *American Economic Review*, 110(5): 1502–39.

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

Ahern, Kenneth R, Ran Duchin, and Tyler Shumway. 2014. "Peer effects in risk aversion and trust." *The Review of Financial Studies*, 27(11): 3213–3240.

Ammermueller, Andreas, and Jörn-Steffen Pischke. 2009. "Peer effects in European primary schools: Evidence from the progress in international reading literacy study." *Journal of Labor Economics*, 27(3): 315–348.

Angrist, Joshua, Peter Hull, and Christopher R Walters. 2022. "Methods for Measuring School Effectiveness."

Argys, Laura M, and Daniel I Rees. 2008. "Searching for peer group effects: A test of the contagion hypothesis." *The Review of Economics and Statistics*, 90(3): 442–458.

Aucejo, Esteban M, Patrick Coate, Jane Fruehwirth, Sean Kelly, and Zachary Mozenter. 2020. "Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations." *Education Policy Analysis Archives*, 28: 62.

Beuermann, Diether W, C Kirabo Jackson, Laia Navarro-Sola, and Francisco Pardo. 2023. "What is a good school, and can parents tell? Evidence on the multidimensionality of school output." *The Review of Economic Studies*, 90(1): 65–101.

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

Bifulco, Robert, Jason M Fletcher, Sun Jung Oh, and Stephen L Ross. 2014. "Do high school peers have persistent effects on college attainment and other life outcomes?" *Labour Economics*, 29: 83–90.

Boisjoly, Johanne, Greg J Duncan, Michael Kremer, Dan M Levy, and Jacque Eccles. 2006. "Empathy or antipathy? The impact of diversity." *American Economic Review*, 96(5): 1890–1905.

Booij, Adam S, Edwin Leuven, and Hessel Oosterbeek. 2017. "Ability peer effects in university: Evidence from a randomized experiment." *The review of economic studies*, 84(2): 547–578.

Boucher, Vincent, Michelle Rendall, Philip Ushchev, and Yves Zenou. 2024. "Toward a general theory

of peer effects." *Econometrica*.

Brady, Ryan R, Michael A Insler, and Ahmed S Rahman. 2017. "Bad Company: Understanding negative peer effects in college achievement." *European Economic Review*, 98: 144–168.

Brunello, Giorgio, Maria De Paola, and Vincenzo Scoppa. 2010. "Peer effects in higher education: Does the field of study matter?" *Economic Inquiry*, 48(3): 621–634.

Burke, Mary A, and Tim R Sass. 2013. "Classroom peer effects and student achievement." *Journal of Labor Economics*, 31(1): 51–82.

Bursztyn, Leonardo, and Robert Jensen. 2015. "How does peer pressure affect educational investments?" *The quarterly journal of economics*, 130(3): 1329–1367.

Bursztyn, Leonardo, Florian Ederer, Bruno Ferman, and Noam Yuchtman. 2014. "Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions." *Econometrica*, 82(4): 1273–1301.

Bursztyn, Leonardo, Georgy Egorov, and Robert Jensen. 2019. "Cool to be smart or smart to be cool? Understanding peer pressure in education." *The Review of Economic Studies*, 86(4): 1487–1526.

Burton, Peter, Shelley Phipps, and Lori Curtis. 2002. "All in the family: A simultaneous model of parenting style and child conduct." *American Economic Review*, 92(2): 368–372.

Bütikofer, Aline, Rita Ginja, Fanny Landaud, and Katrine V Løken. 2020. "School Selectivity, Peers, and Mental Health. IZA Discussion Paper Series, No. 13796." Institute for the Study of Labor (IZA).

Calvó-Armengol, Antoni, Eleonora Patacchini, and Yves Zenou. 2009. "Peer effects and social networks in education." *The Review of Economic Studies*, 76(4): 1239–1267.

Card, David, and Laura Giuliano. 2013. "Peer effects and multiple equilibria in the risky behavior of friends." *Review of Economics and Statistics*, 95(4): 1130–1149.

Carlana, Michela, Eliana La Ferrara, and Paolo Pinotti. Forth.. "Goals and gaps: Educational careers of immigrant children." *Econometrica: Journal of the Econometric Society*.

Carrell, Scott E, and James E West. 2010. "Does professor quality matter? Evidence from random assignment of students to professors." *Journal of Political Economy*, 118(3): 409–432.

Carrell, Scott E, Bruce I Sacerdote, and James E West. 2013. "From natural variation to optimal policy? The importance of endogenous peer group formation." *Econometrica*, 81(3): 855–882.

Carrell, Scott E, Frederick V Malmstrom, and James E West. 2008. "Peer effects in academic cheating." *Journal of human resources*, 43(1): 173–207.

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

Chalendard, Cyril, Ana M Fernandes, Gael Raballand, and Bob Rijkers. 2023. "Corruption in customs." *The Quarterly Journal of Economics*, 138(1): 575–636.

Chang, Simon, Deborah A Cobb-Clark, and Nicolás Salamanca. 2022. "Parents' responses to teacher qualifications." *Journal of Economic Behavior & Organization*, 197: 419–446.

Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *The Quarterly journal of economics*, 126(4): 1593–1660.

Clarke, Damian, Joseph P Romano, and Michael Wolf. 2019. "The Romano-Wolf Multiple Hypothesis Correction in Stata. IZA Discussion Paper Series, No. 12845." Institute for the Study of Labor (IZA).

Cobb-Clark, Deborah A, Nicolas Salamanca, and Anna Zhu. 2019. "Parenting style as an investment in human development." *Journal of Population Economics*, 32(4): 1315–1352.

Correia, S. 2018. "REGHDFE: Stata Module to Perform Linear or Instrumental-Variable Regression Absorbing Any Number of High-Dimensional Fixed Effects. Stat Software Components. Published Online First: September 17, 2018." *Statistical Software Components s457874, Boston College Department of Economics*.

Cullen, Julie Berry, Brian A Jacob, and Steven Levitt. 2006. "The effect of school choice on participants: Evidence from randomized lotteries." *Econometrica*, 74(5): 1191–1230.

Cunha, Flavio. 2015. "Subjective rationality, parenting styles, and investments in children." In *Families in an era of increasing inequality*. 83–94. Springer.

Davezies, Laurent, Guillaume Hollard, and Pedro Vergara Merino. 2024. "Revisiting Randomization with the Cube Method."

de Gendre, Alexandra, Krzysztof Karbownik, Nicolás Salamanca, and Yves Zenou. 2024. "Integrating Minorities in the Classroom: The Role of Students, Parents, and Teachers." National Bureau of Economic Research.

Deming, David J. 2011. "Better schools, less crime?" *The Quarterly Journal of Economics*, 126(4): 2063–2115.

Doepke, Matthias, and Fabrizio Zilibotti. 2017. "Parenting with style: Altruism and paternalism in intergenerational preference transmission." *Econometrica*, 85(5): 1331–1371.

Doepke, Matthias, Giuseppe Sorrenti, and Fabrizio Zilibotti. 2019. "The economics of parenting." *Annual Review of Economics*, 11: 55–84.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya." *American Economic Review*, 101(5): 1739–74.

Elsner, Benjamin, and Ingo E Isphording. 2017. "A big fish in a small pond: Ability rank and human capital investment." *Journal of Labor Economics*, 35(3): 787–828.

Fang, Guanfu, and Shan Wan. 2020. "Peer effects among graduate students: Evidence from China." *China Economic Review*, 60: 101406.

Feld, Jan, and Ulf Zölitz. 2017. "Understanding peer effects: On the nature, estimation, and channels of peer effects." *Journal of Labor Economics*, 35(2): 387–428.

Fredriksson, Peter, Björn Öckert, and Hessel Oosterbeek. 2016. "Parental responses to public investments in children: Evidence from a maximum class size rule." *Journal of Human Resources*, 51(4): 832–868.

Fruehwirth, Jane Cooley. 2014. "Can achievement peer effect estimates inform policy? a view from inside the black box." *Review of Economics and Statistics*, 96(3): 514–523.

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

Gelbach, Jonah B. 2016. "When do covariates matter? And which ones, and how much?" *Journal of Labor Economics*, 34(2): 509–543.

Ghanem, Dalia, and Junjie Zhang. 2014. "'Effortless Perfection:'Do Chinese cities manipulate air pollution data?" *Journal of Environmental Economics and Management*, 68(2): 203–225.

Golsteyn, Bart, Arjan Non, and Ulf Zölitz. Forth.. "The impact of peer personality on academic achievement." *Journal of Political Economy*.

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

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

Hanushek, Eric A, John F Kain, Jacob M Markman, and Steven G Rivkin. 2003. "Does peer ability affect student achievement?" *Journal of applied econometrics*, 18(5): 527–544.

Hao, Lingxin, V Joseph Hotz, and Ginger Z Jin. 2008. "Games parents and adolescents play: Risky behaviour, parental reputation and strategic transfers." *The Economic Journal*, 118(528): 515–555.

Hoekstra, Mark, Pierre Mouganie, and Yaojing Wang. 2018. "Peer quality and the academic benefits to attending better schools." *Journal of Labor Economics*, 36(4): 841–884.

Hoxby, Caroline. 2000. "Peer effects in the classroom: Learning from gender and race variation." National Bureau of Economic Research.

Hoxby, Caroline M, and Gretchen Weingarth. 2005. "Taking race out of the equation: School reassignment and the structure of peer effects." Citeseer.

Imberman, Scott A, Adriana D Kugler, and Bruce I Sacerdote. 2012. "Katrina's children: Evidence on the structure of peer effects from hurricane evacuees." *American Economic Review*, 102(5): 2048–82.

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

Jacob, Brian A, and Steven D Levitt. 2003. "Rotten apples: An investigation of the prevalence and predictors of teacher cheating." *The Quarterly Journal of Economics*, 118(3): 843–877.

Janzen, Sarah A, Nicholas Magnan, Sudhindra Sharma, and William M Thompson. 2017. "Aspirations failure and formation in rural Nepal." *Journal of Economic Behavior & Organization*, 139: 1–25.

Jochmans, Koen. 2023. "Testing random assignment to peer groups." *Journal of Applied Econometrics*, 38(3): 321–333.

Kane, Thomas J, Eric S Taylor, John H Tyler, and Amy L Wooten. 2011. "Identifying effective classroom practices using student achievement data." *Journal of human Resources*, 46(3): 587–613.

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

Kimbrough, Erik O, Andrew McGee, and Hitoshi Shigeoka. Forth.. "How Do Peers Impact Learning? An Experimental Investigation of Peer-to-Peer Teaching and Ability Tracking." *Journal of Human Resources, Forthcoming*.

Kremer, Michael, and Dan Levy. 2008. "Peer effects and alcohol use among college students." *Journal of Economic perspectives*, 22(3): 189–206.

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

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

Lavy, Victor, Olmo Silva, and Felix Weinhardt. 2012. "The good, the bad, and the average: Evidence on ability peer effects in schools." *Journal of Labor Economics*, 30(2): 367–414.

Law, Wing-Wah. 2004. "Translating globalization and democratization into local policy: Educational reform in Hong Kong and Taiwan." *International Review of Education*, 50: 497–524.

Lazear, Edward P. 2001. "Educational production." *The Quarterly Journal of Economics*, 116(3): 777–803.

Le Moglie, Marco, and Giuseppe Sorrenti. 2022. "Revealing "mafia inc."? Financial crisis, organized crime, and the birth of new enterprises." *Review of Economics and Statistics*, 104(1): 142–156.

Lim, Jaegeum, and Jonathan Meer. 2017. "The impact of teacher–student gender matches random assignment evidence from South Korea." *Journal of Human Resources*, 52(4): 979–997.

Lim, Jaegeum, and Jonathan Meer. 2020. "Persistent Effects of Teacher–Student Gender Matches." *Journal of Human Resources*, 55(3): 809–835.

Marmaros, David, and Bruce Sacerdote. 2002. "Peer and social networks in job search." *European economic review*, 46(4-5): 870–879.

Mehta, Nirav, Ralph Stinebrickner, and Todd Stinebrickner. 2019. "Time-Use and Academic Peer Effects in College." *Economic Inquiry*, 57(1): 162–171.

Moretti, Enrico. 2011. "Social learning and peer effects in consumption: Evidence from movie sales." *The Review of Economic Studies*, 78(1): 356–393.

Oliva, Paulina. 2015. "Environmental regulations and corruption: Automobile emissions in Mexico City." *Journal of Political Economy*, 123(3): 686–724.

Oster, Emily, and Rebecca Thornton. 2012. "Determinants of technology adoption: Peer effects in menstrual cup take-up." *Journal of the European Economic Association*, 10(6): 1263–1293.

Pei, Zhuan, Jörn-Steffen Pischke, and Hannes Schwandt. 2019. "Poorly measured confounders are more useful on the left than on the right." *Journal of Business & Economic Statistics*, 37(2): 205–216.

Pop-Eleches, Cristian, and Miguel Urquiola. 2013. "Going to a better school: Effects and behavioral responses." *American Economic Review*, 103(4): 1289–1324.

Romano, Joseph P, and Michael Wolf. 2005*a*. "Exact and approximate stepdown methods for multiple hypothesis testing." *Journal of the American Statistical Association*, 100(469): 94–108.

Romano, Joseph P, and Michael Wolf. 2005*b*. "Stepwise multiple testing as formalized data snooping." *Econometrica*, 73(4): 1237–1282.

Rubin, Mark. 2021. "When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing." *Synthese*, 199(3-4): 10969–11000.

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

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

Stinebrickner, Todd, and Ralph Stinebrickner. 2001. "Peer effects among students from disadvantaged backgrounds." University of Western Ontario, Centre for Human Capital and Productivity (CHCP).

Todd, Petra E, and Kenneth I Wolpin. 2003. "On the specification and estimation of the production function for cognitive achievement." *The Economic Journal*, 113(485): F3–F33.

Todd, Petra E, and Kenneth I Wolpin. 2007. "The production of cognitive achievement in children: Home, school, and racial test score gaps." *Journal of Human capital*, 1(1): 91–136.

Ushchev, Philip, and Yves Zenou. 2020. "Social norms in networks." *Journal of Economic Theory*, 185: 104969.

Vigdor, Jacob, and Thomas Nechyba. 2007. "Peer effects in North Carolina public schools." *Schools and the equal opportunity problem*, 73–101.

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

Xu, Di, Qing Zhang, and Xuehan Zhou. 2020. "The Impact of Low-Ability Peers on Cognitive and Non-Cognitive Outcomes: Random Assignment Evidence on the Effects and Operating Channels." *Journal of Human Resources*.

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

Zárate, Román Andrés. Forth.. "Uncovering Peer Effects in Social and Academic Skills." *American Economic Journal: Applied Economics*.

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

Appendix A The Fishing Algorithm in 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 a large Monte-Carlo simulation of the performance of the algorithm in detecting schools that systematically sort students into classrooms. We cannot provide Monte-Carlo evidence since i) the finite mixture models in Step 3 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; and ii) in Steps 3 and 4 identifying whether a latent class in the model (if any) captures non-compliant risk sets requires some judgement ion whether class posterior means are "close enough" to one, which cannot be automatized (see Section 2.2). 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.

A.1 The Data Generating Process (DGP)

We simulate data in a nested structure and setting that that closely follows our empirical application in Taiwan (see Section 3: students are divided into schools and, within schools, assigned to classrooms. In our simulations 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 elements 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 Notowidigdo (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 soerting test proposed by Jochmans (2023), 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:

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

where $\mathbb{1}[sorting\ school]$ is a dummy variable which flags sorting schools, $p \in [0,1]$ is a *predictor strength* parameter, and $U[0,1]$ is another independent random uniform variable. If p equals 1, $predictor$ will be a perfect determinant of whether a school is systematically sorting students into classrooms; if p equals zero, $predictor$ will be completely uninformative for school type. 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:

$$sorting\ strength = \begin{cases} \theta ability + (1-\theta)U[0,1] & \text{if student is in a sorting school} \\ U[0,1] & \text{otherwise} \end{cases}$$

where $\theta \in [0,1]$ is the parameter that governs the sorting strength in sorting schools and we vary it across DGPs.

The way θ 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 versions of our DGP that correspond to cases of particular interest for an econometrician interested in applying our method:

1. $N_{sorting} = 50; \theta = 0.8; p = 0.8$: 50 strongly sorting schools with a good sorting predictor
2. $N_{sorting} = 50; \theta = 0.8; p = 0.1$: 50 strong sorting schools with a weak sorting predictor

3. $N_{sorting} = 300; \theta = 0.15; p = 0.8$: 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 fail to comply with 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 sorter schools with our method—or any other method we know of for that matter.

A.2 Performance of the 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 as described in Sections 2.2 and 3.3.2, 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 downside 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.

A.3 Case 1: Few Strong Sorter Schools and a Strong Class Predictor

Table A.1 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 Section 2). 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, which is 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

Table A.1: 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 (2023) 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 Model selection on school-level data					
Model BIC for:					
2 latent classes	**317**	**337**	327	**313**	328
3 latent classes	327	344	**322**	321	**327**
4 latent classes	319	351	335	325	330
LR for model with sorting predictor (p-value)	<0.001	-	-	<0.001	-
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%	42%	91%	100%	93%
Pr[Non-sorter school | Defier]	0%	78%	34%	0%	30%
Pr[Sorter school | Complier]	0%	0%	0%	0%	0%
Panel D: Sorting t-statistics in student-level data in classified schools					
Jochmans (2023) 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 percent critical value in the reference a standard normal distribution. In Panel B, numbers in **bold** correspond to the smallest Bayesian Information Criterion (BIC) and the largest predicted outcome mean, used to select the preferred model, and numbers in italics correspond to models that did not comply with convergence criteria. Missing Likelihood Ratio (LR) test p-values in Panel B indicate that either the model using sorting predictors for the latent classes or the model without predictors did not converge (almost always the former).*

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 A.1. 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 in fact have 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 from 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, the better this latent class identifies sorter schools.

Panel C of table A.1 summarizes the performance of the preferred FMM for classifying sorter schools—schools which, in violation of random assignment, systematically sort students into classrooms. We flag sorter 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 Section 2. 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 sorters (out of 300), ii) the percentage of schools that are correctly classified as sorter schools by the Fishing Algorithm and are truly sorter schools, iii) the probability of being wrongly classified as a sorter school and actually being a non-sorter school (false positives), and iv) the probability of being classified as a non-sorter 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)).

In column (2) the Fishing Algorithm performs less well: the algorithm indicates that the majority of schools as sorters, over 50 percent 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 occurred: the selected FMM model in this instance could not use as a latent class predictor to identify the latent class with sorter 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.

Finally, Panel D of Table A.1 shows the sorting test performance from Jochmans (2023) in the simulated data classified as non-sorters by the Fishing Algorithm. For the two models with perfect performance (Columns (1) and (4)), we see that the t-statistics exactly match the non-sorter t-statistics in Panel A, as they should. For the other three models, we see negative and significant t-statistics (Columns (3) and (5)); much more negative for the worst-performing model (Column (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 sorter schools. In Section 2, 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 (2023)'s 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 by changing the number of classes or by exploring alternative 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 Column (1), and Columns (3) to (5) in Panel B.

A.4 Case 2: Few Strong Sorter Schools and a Weak Class Predictor

Table A.2 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.

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 A.1). As a direct result, Panel C shows much higher rates of misclassification, driven entirely by a higher rate of non-sorter schools identified as sorters; 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 non-sorter schools, and positive sorting in the classified sorter schools. In sum, Table A.2 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 (2023) sorting t-statistic for the classified non-sorters subsample. Compared to the findings of Table A.1, the findings of Table A.2 indicate that finding one or multiple strong class predictors is crucial for preventing the algorithm from over-trimming the sample.

Table A.2: Fishing Algorithm Performance in Five Simulated Datasets with 50 Strongly Sorting Schools ($N_{sorting} = 50, \theta = 0.8$) 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 (2023) 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**	**323**	**324**	**309**	**323**
3 latent classes	317	325	331	313	331
4 latent classes	330	337	342	330	346
LR for model with sorting predictor (p-value)	<0.001	<0.001	<0.001	<0.001	-
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	41%	39%	53%	50%	48%
Pr[Non-sorter school | Defier]	78%	79%	74%	75%	76%
Pr[Sorter school | Complier]	0%	0%	0%	0%	0%
Panel D: Sorting t-statistics in student-level data in classified schools					
Jochmans (2023) 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 percent critical value in the reference a standard normal distribution. In Panel B, numbers in **bold** correspond to the smallest Bayesian Information Criterion (BIC) and the largest predicted outcome mean, used to select the preferred model, and numbers in italics correspond to models that did not comply with convergence criteria. Missing Likelihood Ratio (LR) test p-values in Panel B indicate that either the model using sorting predictors for the latent classes or the model without predictors did not converge (almost always the former).*

A.5 Case 3: Weak but Generalized Sorting

Table A.3 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 (2023) t-statistic. This corresponds to setting with generalized non-compliance with 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 percent level. Panel B shows that i) the FMMs in this setting tend to choose 3- and 4-class

Table A.3: 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 (2023) 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	104	85	105	99
3 latent classes	**93**	92	**66**	**103**	95
4 latent classes	95	**85**	70	89	**92**
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%	22%	27%	36%	15%
Pr[Non-sorter school | Defier]	-	-	-	-	-
Pr[Sorter school | Complier]	-	-	-	-	-
Panel D: Sorting t-statistics in student-level data in classified schools					
Jochmans (2023) 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 percent critical value in the reference a standard normal distribution. In Panel B, numbers in **bold** correspond to the smallest Bayesian Information Criterion (BIC) and the largest predicted outcome mean, used to select the preferred model, and numbers in italics correspond to models that did not comply with convergence criteria. Missing Likelihood Ratio (LR) test p-values in Panel B indicate that either the model using sorting predictors for the latent classes or the model without predictors did not converge (almost always the former).*

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 A.2 but lower than in Table A.3. This high predicted sorting strength results in relatively few schools identified as sorters, as show in Panel C. Because the FMMs classify as sorters the schools where the strongest sorting occurs, Panel D again shows strong flipping in the Jochmans (2023) t-statistic.

Overall, Table A.3 indicates that situations where all schools sort students into classrooms (generalized sorting) compared to localized 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 (2023) sorting t-statistic for identified non-sorter schools (Panel D).

A.6 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:

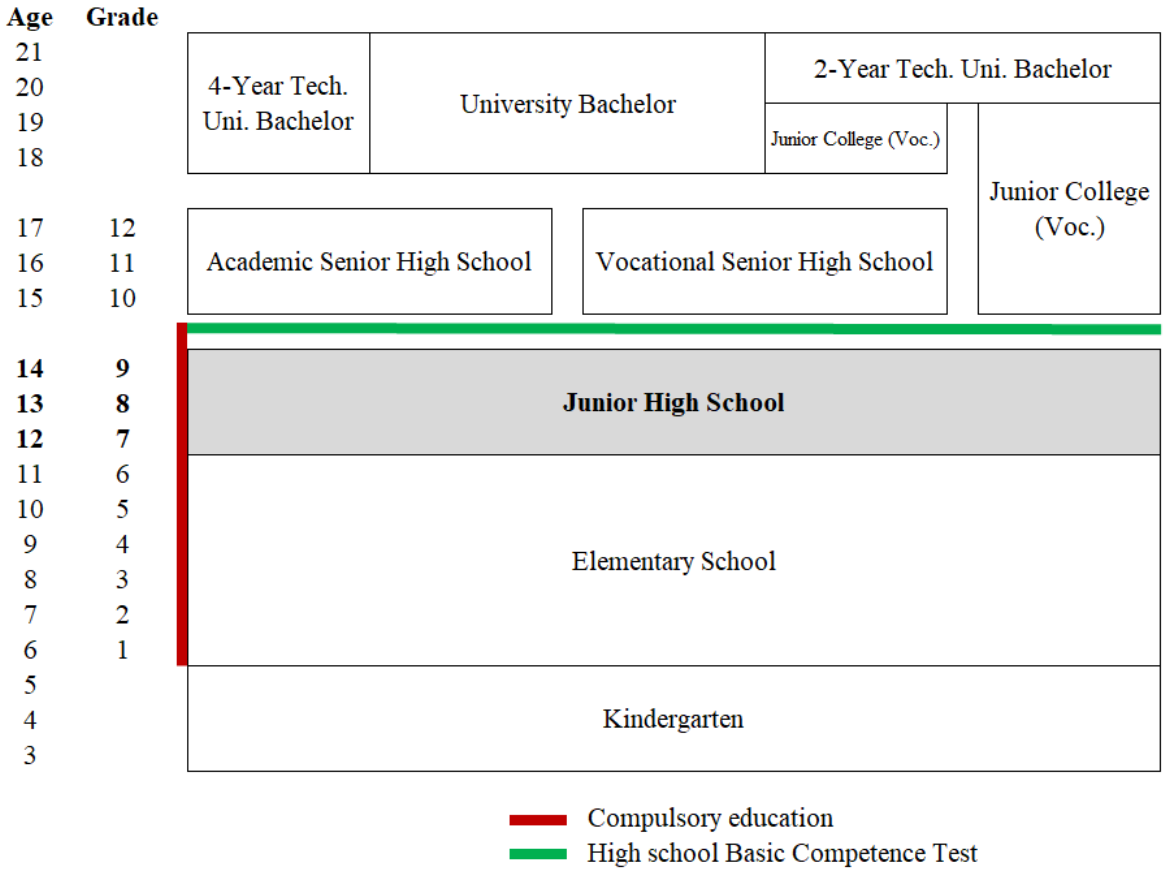
1. Strive to find predictors of whether a school sorts students into classroom, even if these predictors are not perfect. Though technically not necessary, good predictors will meaningfully improve the performance of our method. Place more trust in applications with institutionally sound sorting predictors 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 contains 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 non-sorter schools—is a sign of either over-trimming or widespread sorting.

Appendix B Additional Tables and Figures

Figure B.1: The Education System in Taiwan



This figure shows a simplified schematic of the educational system in Taiwan as it was in place around 2001. The elementary education curriculum covered a range of subjects, including Chinese, mathematics, science, social studies, art, music, and physical education. The junior high school education curriculum continued to cover the same subjects as elementary education, but with greater depth and complexity. In the non-compulsory senior high school education, students could choose to enroll in either general or vocational tracks, with the general track leading to university admission and the vocational track leading to job-specific training. The special education sector, designed for students with special needs such as physical or mental disabilities, is not represented in the figure. Special education was integrated into the mainstream education system as much as possible, but also had specialized schools and facilities for students who required more specialized instruction

Table B.1: Mean of Pre-Assignment and School-Level Variables in the TEPS for the Entire Data and the Trimmed Sample

	Mean of pre-assignment characteristics by sample:	
	TEPS	*Trimmed*
Correct answers on standardized test	40.9	40.7
Female student	0.50	0.48
Student year of birth	1988.6	1988.6
No. of siblings of student	1.77	1.77
Responding parent is female	0.64	0.63
Ethnic minority father	0.05	0.04
Two-parent household	0.86	0.87
Father's birth year	1958.6	1958.7
Father has post-secondary education	0.12	0.11
Unemployed father	0.11	0.11
Household monthly income is		
NT$20,000 or less	0.11	0.10
NT$20,000-NT$50,000	0.41	0.41
NT$50,000-NT$100,000	0.35	0.35
More than NT$100,000	0.14	0.13
Classroom size	35.9	36.4
Male-to-female students ratio	0.52	0.52
Number of sampled students in school	67.0	67.5
School size	479	492
School sampling rate	0.19	0.18
No. of students (approx.)	20,055	12,793

This table presents summary statistics of student and parent demographic characteristics in the complete TEPS data which includes of up to 333 schools and 1,244 classrooms and 20,055 students, and in our estimation sample trimmed by the Fishing Algorithm which includes of up to 231 schools and 851 classrooms and 12,793 students.

Table B.2: Brief Description of All Relevant Wave 2 Measures in the TEPS Data

Measure	Description	Items \ Values
	Students	
Test scores	Comprehensive Analytical Ability standardized test, measure of cognitive ability	75\59
Study time	Hours of study in and out of school, during private tutoring, other academic activities, study ethos, summer activities	7 \ 25
Truant behavior	Ever skipped class, got into fights, smoked tobacco or drank alcohol, watched porn, ran away from home, stole	1 \ 2
Cheated on exams	Ever cheated on exams	1 \ 2
Academic self-efficacy	Focus, diligence, conscientiousness, initiative, eloquence, organization, cooperation, curiosity	10 \ 19
Mental health	Feeling troubled, depressed, suicidal, nervous, unfocused, pressured, irritated, isolated, guilty	12 \ 22
University aspirations	Student wants to go to university	1 \ 2
University expectations	Student expects to be able to go to university	1 \ 2
	Parents	
Money investments	Out-of-school tutoring for student: cost and intensity	3 \ 10
Time investments	Frequency of going to bookstores and cultural events together with student	2 \ 11
Parent-child conflict	Student quarrels with father and mother	1 \ 2
Parental strictness	Father and mother's strict discipline with student	2 \ 17
Parental support	Father and mother discuss future, listen carefully, worry and give advice, accept student unconditionally	8 \ 7
Harsh parenting	Parents use harsh punishment with student	1 \ 2
University aspirations	Parents want student to go to university	1 \ 2
	Schools & Teachers	
School environment	Student perception of school study ethos, campus safety, school fairness, engagement of school administrators	5 \ 16
Classroom hard to manage	Teacher reports how hard to manage this classroom	1 \ 2
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 \ 36
Teacher tired of teaching	Teacher reports how tired of teaching	1 \ 2

This table shows the number of items measuring student test scores, and educational inputs of students, parents and teachers in Wave 2 of the TEPS. After checking for internal consistency and performance, we combine the items of most of these measures into summative scales. Truant behavior, cheating on exams, students' university aspirations and expectations, parent-child conflict, whether classrooms are hard to manage and whether teachers are tired are measured using dummy variables. The table reports the total values taken by each of these measures.

Table B.3: The Effect of Peer Test Scores in Wave 1 on Students' Own Test Scores in Wave 2

Outcome:	*Student test scores in wave 2 [std]*			
Peer test scores [std]	0.046***	0.048***	0.043***	0.037**
	(0.017)	(0.017)	(0.016)	(0.016)
W1 test scores	✓	✓	✓	✓
Balancing controls		✓	✓	✓
All pre-assignment			✓	✓
Other peer char.				✓
R^2	0.64	0.65	0.65	0.65
Schools	231	231	231	231
Classes	851	851	851	851
Students	12,793	12,793	12,793	12,793

*This table shows coefficient estimates of regressing standardized student test scores in wave 2 on standardized average peer test scores in wave 1 in our trimmed sample of up to 231 schools and 851 classrooms and 12,793 students. Estimates in this table are also shown in Figure 3. W1 test scores stand for student's own test scores in wave 1. Balancing controls stands for pre-assignment control characteristics unbalanced in wave 1 (household income, family help with homework and whether parents tried to influence junior high classroom assignment). All preassignment stands for all 17 pre-assignment characteristics tested for balancing. Other peer char. stands for the leave-out mean of other peer characteristics including gender, age, family income and education, government-employed parents, ethnic minority status, and reports of attending gifted classes, or having pushy parents. All models include school fixed effects. Missing covariates are imputed at the median and a missing covariate flag is always added. Standard errors are clustered at the classroom level, and coefficients statistically different from zero at the 99, 95 and 90 percent confidence level are marked with ***, **, and *.*

Table B.4: The Effect of Peer Test Scores in Wave 1 on Educational Inputs in Wave 2

Treatment:		*Peer test scores [std]*					
	Mean	*Coef.*	*Std. err.*	R^2	*Schools*	*Classes*	*Students*
Outcomes: educational inputs							
of students							
Study time	std	-0.027	(0.023)	0.16	231	851	12,758
Truant behavior	0.39	0.003	(0.011)	0.17	231	851	12,733
Cheated on exams	0.49	0.012	(0.015)	0.09	231	851	12,690
Academic self-efficacy	std	-0.032	(0.021)	0.08	231	851	12,737
Mental health	std	-0.021	(0.023)	0.07	231	851	12,729
University aspirations	0.53	0.012	(0.010)	0.20	231	851	12,735
University expectations	0.41	0.009	(0.010)	0.21	231	851	12,726
of parents							
Private tutoring	std	0.011	(0.021)	0.24	231	851	12,789
Time with parents	std	0.036*	(0.020)	0.08	231	851	12,719
Conflict with parents	0.30	-0.015	(0.010)	0.08	231	851	12,689
Parental strictness	std	0.020	(0.022)	0.07	231	851	12,757
Parental support	std	0.006	(0.022)	0.09	231	851	12,757
Harsh parenting	0.33	0.021**	(0.010)	0.04	231	851	12,757
Parent uni. aspirations	0.47	-0.006	(0.011)	0.25	231	851	12,618
of schools							
School environment	std	-0.046*	(0.027)	0.10	231	851	12,748
Classroom hard to manage	0.32	-0.090**	(0.038)	0.28	231	836	11,967
Teacher engagement	std	-0.001	(0.025)	0.07	231	851	12,752
Teacher tired of teaching	0.49	-0.038	(0.037)	0.29	231	836	11,964

*This table shows coefficient estimates of standardized average peer test scores in wave 1 on measures of student, parent, and teacher educational inputs between waves 1 and 2 (measured in wave 2) in our trimmed sample of up to 231 schools and 851 classrooms and 12,793 students. Estimates in this table are also shown in Figure 4. Different rows correspond to different educational inputs. Inputs measured as standardized indices (with a mean of zero and a standard deviation of one) are marked as 'std', else the unconditional mean of binary inputs is reported. All models control for student's own test scores in wave 1, all 17 pre-assignment characteristics tested for balancing, the leave-out mean of peer characteristics, and school fixed effects. Missing covariates are imputed at the median and a missing covariate flag is always added. Standard errors are clustered at the classroom level, and coefficients statistically different from zero at the 99, 95 and 90 percent confidence level are marked with ***, **, and *.*

Table B.5: Permutation-Based Balancing Tests of Peer Test Scores

	Share of classes with empirical p-values under			
	0.10	*0.05*	*0.01*	*Avg. p-value*
Pre-assignment characteristics				
Student test scores	0.11	0.05	0.02	0.483
Female student	0.07	0.03	0.02	0.561
Student born before 1989	0.11	0.06	0.01	0.499
Monthly household income over NT$100,000	0.09	0.04	0.01	0.497
College-educated parent(s)	0.09	0.05	0.02	0.489
Parent(s) work in government	0.07	0.04	0.01	0.496
Ethnic minority parent(s)	0.08	0.04	0.02	0.501
Student prioritized studies since primary school	0.11	0.07	0.01	0.484
Student reviews lessons since primary school	0.13	0.07	0.01	0.475
Student likes new things since primary school	0.14	0.08	0.02	0.470
Student was truant in primary school	0.09	0.03	0.01	0.496
Student had mental health issues in primary school	0.11	0.06	0.01	0.498
Had private tutoring before junior high school	0.11	0.06	0.01	0.482
Family help with homework before junior high school	0.09	0.05	0.01	0.502
Student quarreled with parents in primary school	0.10	0.04	0.01	0.503
Student enrolled in gifted academic class	0.09	0.05	0.01	0.482
Student enrolled in arts gifted class	0.11	0.07	0.02	0.459
Parents made efforts to place student in better class	0.13	0.08	0.02	0.479

This table shows the results of permutation-based classroom-level balancing tests in our trimmed sample of up to 231 schools, 851 classrooms and 12,793 students. For these tests, we produce 1,000 simulations of randomly assigned students to classrooms within schools and calculate the mean of pre-assignment characteristics in each synthetic classroom. We then construct classroom-level empirical p-values for each pre-assignment characteristic as the share of the 1,000 simulations where the synthetic classroom mean are more extreme than the actual classroom means. The leftmost three columns show the share of classrooms where the empirical p-values exceed 0.10, 0.05 and 0.01. The rightmost column shows the average empirical p-value across all classrooms.

Table B.6: Non-Parametric Balancing Tests of Peer Test Scores

	Share of classroom-dummy joint significance test p-values under		
	.10	*.05*	*.01*
Outcomes: Pre-assignment characteristics			
Student test scores	0.08	0.04	0.03
Female student	0.04	0.03	0.01
Student born before 1989	0.13	0.07	0.01
Monthly household income over NT$100,000	0.08	0.04	0.00
College-educated parent(s)	0.12	0.07	0.03
Parent(s) work in government	0.06	0.04	0.01
Ethnic minority parent(s)	0.05	0.04	0.01
Student prioritized studies since primary school	0.12	0.08	0.01
Student reviews lessons since primary school	0.14	0.07	0.03
Student likes new things since primary school	0.16	0.10	0.02
Student was truant in primary school	0.08	0.04	0.00
Student had mental health issues in primary school	0.11	0.05	0.01
Had private tutoring before junior high school	0.13	0.07	0.02
Family help with homework before junior high school	0.08	0.04	0.02
Student quarreled with parents in primary school	0.09	0.04	0.00
Student enrolled in gifted academic class	0.10	0.04	0.02
Student enrolled in arts gifted class	0.15	0.11	0.07
Parents made efforts to place student in better class	0.16	0.10	0.04

This table shows the results of non-parametric school-level balancing tests in our trimmed sample of of up to 231 schools, 851 classrooms and 12,793 students. School-by-school, we regress each pre-assignment characteristic on a set of classroom dummies, and F-test them for joint significance. The three leftmost columns show the share of schools for which the p-values fall under 0.10, 0.05 and 0.01. Each row presents these for a different pre-assignment characteristic. We use cluster-robust covariance matrices at the classroom level for each test.

Table B.7: Sensitivity of Main Estimates to Different Latent Sorter Probability Thresholds

Treatment:	*Peer test scores [std]*				
Exclude school if Pr[Defier]<	*0.2*	*0.4*	*0.6*	*0.8*	*1*
Outcomes:					
Test scores	0.038**	0.043**	0.035**	0.030**	0.055***
	(0.017)	(0.017)	(0.016)	(0.014)	(0.011)
Study time	-0.015	-0.026	-0.031	-0.025	0.016
	(0.024)	(0.024)	(0.023)	(0.019)	(0.015)
Truant behavior	0.001	-0.002	0.001	0.006	-0.013*
	(0.012)	(0.012)	(0.011)	(0.009)	(0.007)
Cheated on exams	0.012	0.007	0.015	0.028**	0.014
	(0.016)	(0.015)	(0.015)	(0.012)	(0.009)
Academic self-efficacy	-0.028	-0.032	-0.041*	-0.025	-0.014
	(0.023)	(0.022)	(0.021)	(0.018)	(0.014)
Mental health	-0.018	-0.025	-0.022	-0.025	0.003
	(0.024)	(0.023)	(0.022)	(0.018)	(0.014)
University aspirations	0.014	0.014	0.011	0.007	0.007
	(0.011)	(0.010)	(0.010)	(0.008)	(0.006)
University expectations	0.009	0.010	0.009	0.004	0.008
	(0.011)	(0.010)	(0.010)	(0.009)	(0.006)
Private tutoring	0.026	0.022	0.013	0.032*	0.013
	(0.023)	(0.021)	(0.021)	(0.018)	(0.013)
Time with parents	0.050**	0.039*	0.033*	0.025	0.016
	(0.021)	(0.021)	(0.020)	(0.017)	(0.015)
Conflict with parents	-0.009	-0.012	-0.014	-0.008	-0.008
	(0.011)	(0.011)	(0.010)	(0.008)	(0.006)
Parental strictness	0.025	0.012	0.014	-0.014	0.005
	(0.023)	(0.023)	(0.022)	(0.018)	(0.014)
Parental support	0.008	0.008	0.005	0.012	0.027**
	(0.022)	(0.022)	(0.021)	(0.018)	(0.013)
Harsh parenting	0.025**	0.023**	0.021**	0.008	0.004
	(0.011)	(0.010)	(0.010)	(0.008)	(0.006)
Parent uni. aspirations	-0.006	-0.004	-0.006	0.002	0.003
	(0.011)	(0.011)	(0.010)	(0.008)	(0.006)
School environment	-0.064**	-0.049*	-0.044	-0.029	0.018
	(0.028)	(0.028)	(0.027)	(0.021)	(0.016)
Classroom hard to manage	-0.097**	-0.100**	-0.087**	-0.088***	-0.099***
	(0.040)	(0.039)	(0.037)	(0.030)	(0.022)
Teacher engagement	-0.010	-0.003	-0.006	-0.007	-0.005
	(0.027)	(0.026)	(0.025)	(0.020)	(0.016)
Teacher tired of teaching	-0.071*	-0.049	-0.035	-0.011	-0.014
	(0.039)	(0.038)	(0.036)	(0.030)	(0.023)
Jochman's t-statistic	-1.4	-1.1	-0.7	**2.6**	**6.6**
Schools	216	224	233	267	333
Classrooms	803	828	859	995	1,244
Students	12,099	12,454	12,907	14,808	18,655

*This table shows coefficient estimates of regressing student outcomes in wave 2 on standardized average peer ability in wave 1 in the complete TEPS which includes of up to 333 schools and 1,244 classrooms and 20,055 students. Estimation samples samples defined by taking different thresholds, τ, in the school-level posterior probability of being a defier school, as defined by the Fishing Algorithm. All models include school fixed effects, and students' own ability and educational inputs in wave 1. At the bottom we report the Jochmans (2023) sorting t-statistics, noting that their reference distribution is the standard normal. T-statistics larger than critical values for a two-sided test are shown in italics for 95 percent confidence and in bold for 99 percent confidence. Standard errors are clustered at the classroom level, and coefficients statistically different from zero at the 99, 95 and 90 percent confidence level are marked with ***, **, and *.*

Table B.8: Sorting and Balancing Tests Using the Complete TEPS Data Weighted by $1 - \hat{P}_{sl}$

Panel A: Sorting tests on test scores				
	Students	*Mean*	*Sorting test t-statistic*	
Guryan et al. (2009)	19,957	Std	0.83	
Jochmans (2023)	19,957	Std	-0.69	
Panel B: Balancing tests on pre-assignment characteristics				
	Students	*Mean*	*Peer test scores*	
			Coef.	*Std. err.*
Female student	19,957	0.48	0.009	(0.009)
Student born before 1989	19,866	0.37	-0.010	(0.008)
Household income > NT$100k/mo.	19,629	0.14	-0.009*	(0.005)
College-educated parent(s)	19,073	0.13	0.010	(0.007)
Parent(s) work in government	18,979	0.09	0.013**	(0.005)
Ethnic minority parent(s)	19,070	0.06	-0.007	(0.008)
Prioritized studies since primary school	19,830	0.26	-0.006	(0.007)
Reviews lessons since primary school	19,813	0.17	0.003	(0.007)
Likes new things since primary school	19,771	0.41	0.001	(0.009)
Was truant in primary school	19,674	0.34	-0.002	(0.009)
Student had mental health issues in primary school	19,670	0.47	-0.000	(0.008)
Had private tutoring before junior high	19,720	0.67	0.010	(0.009)
Family help with homework before junior high	18,976	0.83	-0.015**	(0.006)
Student quarreled with parents in primary school	19,691	0.67	-0.004	(0.008)
Student enrolled in gifted academic classroom	19,779	0.06	0.021***	(0.007)
Student enrolled in arts gifted classroom	19,779	0.05	-0.006	(0.013)
Parents made efforts to place student in better classroom	19,698	0.15	0.040***	(0.008)

*This table shows results of sorting and balancing tests on peer test scores in the complete TEPS data which includes up to 333 schools and 1,244 classrooms and 20,055 students. All estimators include school fixed effects and use as analytical weights $1 - \hat{P}_{sl}$, the estimated school-level probability that a school belongs to the latent class of compliers with random assignment (normalized to sum to one). In Panel A, the reference distribution for the Guryan, Kroft and Notowidigdo (2009) and the Jochmans (2023) sorting statistics is the standard normal. In Panel B, the rightmost column reports cluster-robust standard errors at the classroom level. ***, ** and * mark estimates statistically different from zero at the 99, 95 and 90 percent confidence level.*

Table B.9: The Effect of Peer Test Scores in Wave 1 on Student Outcomes in Wave 2 Using the Complete TEPS Data Weighted by $1 - \hat{P}_{sl}$

Treatment:		*Peer test scores [std]*					
	Mean	*Coef.*	*Std. err.*	R^2	*Schools*	*Classes*	*Students*
Outcomes:							
Test scores	Std	0.041***	(0.014)	0.65	333	1,244	18,655
Study time	Std	-0.013	(0.019)	0.16	333	1244	18803
Truant behavior	0.39	-0.000	(0.009)	0.18	333	1244	18769
Cheated on exams	0.49	0.017	(0.012)	0.09	333	1244	18706
Academic self-efficacy	Std	-0.026	(0.018)	0.08	333	1244	18765
Mental health	Std	-0.016	(0.019)	0.07	333	1244	18752
University aspirations	0.53	0.010	(0.008)	0.21	333	1244	18765
University expectations	0.41	0.007	(0.008)	0.21	333	1244	18752
Private tutoring	Std	0.020	(0.018)	0.25	333	1,244	18,885
Time with parents	Std	0.033**	(0.017)	0.07	333	1,244	18,789
Conflict with parents	0.3	-0.009	(0.008)	0.08	333	1,244	18,724
Parental strictness	Std	0.004	(0.018)	0.07	333	1,244	18,801
Parental support	Std	0.011	(0.017)	0.09	333	1,244	18,801
Harsh parenting	0.33	0.015*	(0.008)	0.04	333	1,244	18,801
Parent uni. aspirations	0.47	0.000	(0.009)	0.25	333	1,244	18,639
School environment	Std	-0.033	(0.022)	0.10	333	1,244	18,784
Classroom hard to manage	0.33	-0.092***	(0.031)	0.28	333	1,226	17,772
Teacher engagement	Std	-0.005	(0.021)	0.07	333	1,244	18,790
Teacher tired of teaching	0.49	-0.030	(0.030)	0.28	333	1,224	17,714

This table shows coefficient estimates of the standardized average peer test scores in wave 1 on student test scores in wave 2 and measures of student, parent, and teacher educational inputs between waves 1 and 2 (measured in wave 2) in the complete TEPS data which includes up to 333 schools and 1,244 classrooms and 20,055 students. Different rows correspond to different outcomes. Outcomes measured as standardized indices (with a mean of zero and a standard deviation of one) are marked as 'std', else the unconditional mean of binary outcomes is reported. All models control for student's own test scores in wave 1, all 17 pre-assignment characteristics tested for balancing, the leave-out mean of peer characteristics, and school fixed effects. Missing covariates are imputed at the median and a missing covariate flag is always added. All estimators include school fixed effects and use as analytical weights $1 - \hat{P}_{sl}$*, the estimated school-level probability that a school belongs to the latent class of compliers with random assignment (normalized to sum to one). The rightmost column reports cluster-robust standard errors at the classroom level. ***, ** and * mark estimates statistically different from zero at the 99, 95 and 90 percent confidence level.*

Table B.10: The Effect of Peer Test Scores in Wave 1 on Students' Own Test Scores in Wave 2 Using Alternative Measures of Ability

Outcome:	*Student ability in wave 2 [std] with ability measured as:*				
			IRT Bayesian posterior mean of:		
	Analytical	*Mathematical*	*General*	*Analytical*	*Mathematical*
Peer ability [std]	0.011	0.029*	0.027	0.014	0.026
	(0.019)	(0.016)	(0.017)	(0.019)	(0.017)
R^2	0.42	0.59	0.68	0.44	0.61
Schools	231	231	231	231	231
Classes	851	851	851	851	851
Students	12,793	12,793	12,793	12,793	12,793

*This table shows coefficient estimates of regressing student's own ability in wave 2 on standardized average peer ability in wave 1 in our trimmed sample of up to 231 schools and 851 classrooms and 12,793 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 control for student's own test scores in wave 1, all 17 pre-assignment characteristics tested for balancing, the leave-one-out mean of peer characteristics, and school fixed effects. Missing covariates are imputed at the median and a missing covariate flag is always added. Standard errors are clustered at the classroom level, and coefficients statistically different from zero at the 99, 95 and 90 percent confidence level are marked with ***, **, and *.*

Table B.11: The Effect of Peer Test Scores in Wave 1 on Student Outcomes in Wave 2 Using the Correction for Incomplete Classroom Sampling from Sojourner (2013)

	Effect of peer test scores [std] with Sojourner (2013) correction for peer test scores missing not at random using:					
Share of peers observed						
× School FE	✓					
× School k-tile FE, k =		*25*	*20*	*15*	*10*	*5*
Outcomes:						
Test scores	0.111***	0.082**	0.080**	0.075**	0.084**	0.085**
	(0.039)	(0.037)	(0.037)	(0.036)	(0.036)	(0.036)
Study time	0.003	-0.082*	-0.083*	-0.080	-0.078	-0.078
	(0.054)	(0.050)	(0.049)	(0.049)	(0.048)	(0.049)
Truant behavior	-0.025	-0.017	-0.016	-0.019	-0.011	-0.014
	(0.026)	(0.024)	(0.024)	(0.024)	(0.023)	(0.023)
Cheated on exams	0.026	0.006	0.007	0.016	0.012	0.008
	(0.037)	(0.032)	(0.032)	(0.031)	(0.031)	(0.031)
Academic self-efficacy	-0.085	-0.071	-0.057	-0.056	-0.045	-0.042
	(0.053)	(0.048)	(0.047)	(0.048)	(0.047)	(0.046)
Mental health	-0.039	0.011	0.012	0.013	0.024	0.028
	(0.063)	(0.054)	(0.055)	(0.054)	(0.053)	(0.053)
University aspirations	0.035	0.025	0.032	0.027	0.028	0.025
	(0.026)	(0.022)	(0.023)	(0.023)	(0.023)	(0.023)
University expectations	0.031	0.011	0.010	0.008	0.011	0.012
	(0.023)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Private tutoring	0.016	0.015	0.008	0.024	0.037	0.028
	(0.046)	(0.043)	(0.043)	(0.043)	(0.043)	(0.042)
Time with parents	0.022	0.079*	0.079*	0.063	0.105**	0.097**
	(0.048)	(0.042)	(0.042)	(0.041)	(0.042)	(0.042)
Conflict with parents	0.001	-0.014	-0.009	-0.013	-0.017	-0.016
	(0.024)	(0.021)	(0.021)	(0.021)	(0.020)	(0.020)
Parental strictness	-0.039	0.015	0.029	0.023	0.029	0.024
	(0.052)	(0.048)	(0.047)	(0.048)	(0.046)	(0.046)
Parental support	-0.006	0.010	0.021	0.018	0.022	0.014
	(0.051)	(0.045)	(0.046)	(0.045)	(0.045)	(0.045)
Harsh parenting	0.034	0.022	0.023	0.033	0.028	0.025
	(0.024)	(0.022)	(0.022)	(0.022)	(0.021)	(0.021)
Parent uni. aspirations	0.019	0.001	-0.000	-0.004	-0.004	-0.005
	(0.025)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)
School environment	0.019	-0.031	-0.038	-0.024	-0.049	-0.038
	(0.072)	(0.062)	(0.061)	(0.061)	(0.060)	(0.059)
Classroom hard to manage	-0.300***	-0.237***	-0.232***	-0.242***	-0.234***	-0.236***
	(0.088)	(0.081)	(0.081)	(0.082)	(0.080)	(0.079)
Teacher engagement	0.050	-0.012	-0.001	-0.018	-0.003	0.000
	(0.065)	(0.057)	(0.057)	(0.056)	(0.056)	(0.056)
Teacher tired of teaching	-0.106	-0.065	-0.059	-0.082	-0.093	-0.085
	(0.098)	(0.083)	(0.082)	(0.084)	(0.082)	(0.083)

This table shows coefficient estimates of regressing student outcomes in wave 2 on standardized average peer ability in wave 1 in our trimmed sample of up to 231 schools and 851 classrooms and 12,793 students. These estimates correct for peer test scores missing not at random following Sojourner (2013) and implemented using the Correia (2018) `reghdfe` *Stata package. All models control for student's own test scores in wave 1, all 17 pre-assignment characteristics tested for balancing, the leave-one-out mean of peer characteristics, and school fixed effects. Missing covariates are imputed at the median and a missing covariate flag is always added. Standard errors are clustered at the classroom level, and coefficients statistically different from zero at the 99, 95 and 90 percent confidence level are marked with ***, **, and *.*

Table B.12: *p*-values of the Effects of Peer Test Scores in Wave 1 on Student Outcomes in Wave 2 Under Randomization Inference and After Correcting for Multiple Hypothesis Testing

	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:		
Student test scores	0.042	
Study time	0.327	0.941
Truant behavior	0.788	0.985
Cheated on exams	0.497	0.962
Academic self-efficacy	0.208	0.818
Mental health	0.439	0.962
University aspirations	0.293	0.941
University expectations	0.450	0.962
Private tutoring	0.677	0.981
Time with parents	0.123	0.687
Conflict with parents	0.209	0.818
Parental strictness	0.446	0.962
Parental support	0.812	0.985
Harsh parenting	0.069	0.433
Parent uni. aspirations	0.607	0.941
School environment	0.130	0.699
Classroom hard to manage	0.043	0.261
Teacher engagement	0.966	0.985
Teacher tired of teaching	0.351	0.946

*This table corrected p-values for our main results in our trimmed sample of up to 231 schools and 851 classrooms and 12,793 students using i) the randomization-t inference procedure from Young (2019) to account for high-leverage, finite sample properties of the model error term, and the complex sampling structure of our data (based on 999 permutations), and ii) the step-down procedure from Romano and Wolf (2005*b*) to control for family-wise error rate in multiple hypotheses testing implemented using the* `rwolf` *Stata package from Clarke, Romano and Wolf (2019). Both procedures are based on 999 replications. All models control for student's own test scores in wave 1, all 17 pre-assignment characteristics tested for balancing, the leave-one-out mean of peer characteristics, and school fixed effects. Missing covariates are imputed at the median and a missing covariate flag is always added.*

Table B.13: The Heterogeneous Effects of Peer Test Scores in Wave 1 on Students' Own Test Scores in Wave 2

Outcome:	*Student test scores in wave 2 [std]*		
by student test scores:	*below median*	*above median*	Diff. (p-value)
Peer test scores [std]	0.036**	0.035*	0.967
	(0.018)	(0.021)	
by peer test scores:	*below median*	*above median*	Diff. (p-value)
Peer test scores [std]	0.069***	0.032	0.375
	(0.025)	(0.034)	
by student gender:	*male*	*female*	Diff. (p-value)
Peer test scores [std]	0.034*	0.037**	0.892
	(0.019)	(0.019)	
by household monthly income:	*<NT$50k*	*>NT$50k*	Diff. (p-value)
Peer test scores [std]	0.047**	0.031	0.434
	(0.019)	(0.020)	
by parent(s) college degree:	*No*	*Yes*	Diff. (p-value)
Peer test scores [std]	0.038**	0.026	0.671
	(0.017)	(0.031)	
by Dao Shi experience:	*<10 years*	*>10 years*	Diff. (p-value)
Peer test scores [std]	0.034	0.028	0.846
	(0.027)	(0.018)	

This table shows average marginal effect (AME) estimates of regressing standardized student test scores in wave 2 on standardized average peer test scores in wave 1 for subgroups in our trimmed sample of up to 231 schools and 851 classrooms and 12,793 students. Each estimate is based on a fully interacted regression of the variable defining the subgroup with all our regressors and is calculated using Stata's `margins` *command. Rows present the peer effects for different subgroups defined based on wave 1 variables. All models control for student's own test scores in wave 1, all 17 pre-assignment characteristics tested for balancing, the leave-one-out mean of peer characteristics, and school fixed effects. Missing covariates are imputed at the median and a missing covariate flag is always added. Standard errors are clustered at the classroom level, and coefficients statistically different from zero at the 99, 95 and 90 percent confidence level are marked with ***, **, and *.*

Table B.14: The Effect of Peer Test Scores in Wave 1 on Students' Own Test Scores in High School, and on Educational Achievement and Labor Market Outcomes in Early Adulthood

Treatment:		*Peer test scores [std]*					
	Mean	*Coef.*	*Std. err.*	R^2	*Schools*	*Classes*	*Students*
Outcomes:							
during high school							
Test scores in grade 11 [Std]	std	-0.006	(0.036)	0.68	230	792	2,778
Test scores in grade 12 [Std]	std	-0.044	(0.039)	0.63	230	785	2,679
around age 20							
University (attending or finished)	0.44	0.041*	(0.025)	0.47	228	752	2,115
Vocational education (attending or finished)	0.35	-0.062**	(0.028)	0.30	228	752	2,115
around age 24							
Postgraduate studies (attending or finished)	0.17	0.009	(0.021)	0.26	228	738	2,103
Has a full time job	0.69	0.025	(0.027)	0.29	228	737	2,092
Earns above-median income	0.44	-0.053	(0.040)	0.30	223	665	1,410

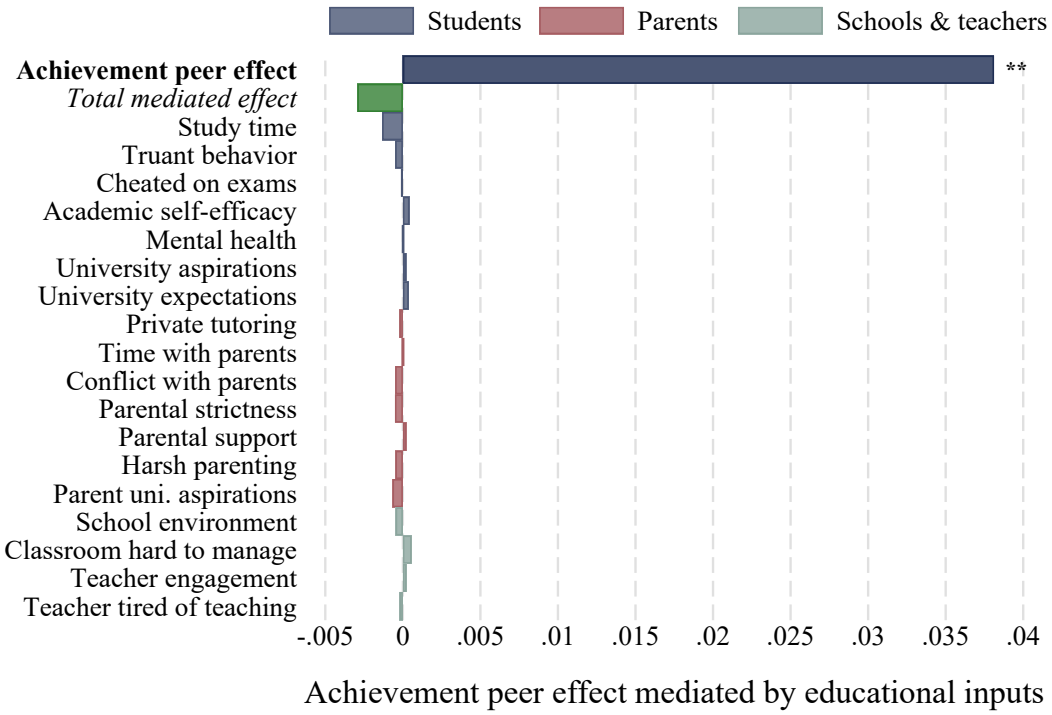
*This table shows estimates of coefficient estimates of standardized average peer test scores in wave 1 on long-term student outcomes measured in waves 3 and 4 of TEPS and through the TEPS Beyond project. These waves sample a non-representative subset of students in TEPS waves 1 and 2. Different rows correspond to different outcomes. Inverse probability weights are used throughout, and these are constructed as the predicted values from linear probability models of sample attrition dummies on own and peer characteristics (windsorized at the 5^{th} and 95^{th} percentile), all preassignment controls, and school fixed effects. All models control for student's own test scores in wave 1, all 17 pre-assignment characteristics tested for balancing, the leave-one-out mean of peer characteristics, and school fixed effects. Missing covariates are imputed at the median and a missing covariate flag is always added. Standard errors are clustered at the classroom level, and coefficients statistically different from zero at the 99, 95 and 90 percent confidence level are marked with ***, **, and *.*

Table B.15: The Estimated Return to Educational Inputs from Different Value-Added Models

Outcome:	*Student test scores in wave 2 [std]*						
Model:	*Contemp.*	*Cumulative*	*School FE*	*Child FE*	*VA*	*Cumulative VA*	*CVA w/ school FE*
Own test scores in W1					0.624	0.604	0.606
					(0.008)	(0.008)	(0.008)
Peer test scores in W1		0.180	-0.024	-0.025		0.052	0.041
		(0.014)	(0.024)	(0.014)		(0.013)	(0.017)
Study time	0.082	0.076	0.073	0.006	0.048	0.047	0.046
	(0.009)	(0.009)	(0.009)	(0.005)	(0.007)	(0.007)	(0.007)
Truant behavior	-0.131	-0.120	-0.115	-0.043	-0.066	-0.068	-0.063
	(0.018)	(0.017)	(0.017)	(0.013)	(0.014)	(0.013)	(0.013)
Cheated on exams	-0.043	-0.030	-0.037	-0.024	-0.018	-0.017	-0.024
	(0.016)	(0.015)	(0.015)	(0.012)	(0.012)	(0.012)	(0.012)
Academic self-efficacy	-0.027	-0.031	-0.029	-0.008	-0.016	-0.018	-0.017
	(0.008)	(0.008)	(0.008)	(0.006)	(0.006)	(0.007)	(0.006)
Mental health	-0.058	-0.053	-0.050	0.002	-0.019	-0.017	-0.015
	(0.008)	(0.008)	(0.008)	(0.006)	(0.006)	(0.007)	(0.007)
University aspirations	0.201	0.150	0.149	0.026	0.088	0.075	0.072
	(0.019)	(0.018)	(0.018)	(0.011)	(0.014)	(0.014)	(0.014)
University expectations	0.381	0.328	0.322	0.028	0.187	0.172	0.165
	(0.019)	(0.019)	(0.018)	(0.012)	(0.015)	(0.015)	(0.014)
Private tutoring	0.110	0.084	0.085	0.008	0.041	0.036	0.039
	(0.009)	(0.009)	(0.009)	(0.006)	(0.007)	(0.007)	(0.007)
Time with parents	-0.016	-0.013	-0.006	-0.000	-0.005	-0.005	0.000
	(0.008)	(0.008)	(0.008)	(0.005)	(0.006)	(0.006)	(0.006)
Conflict with parents	0.093	0.091	0.090	0.029	0.055	0.057	0.058
	(0.017)	(0.016)	(0.016)	(0.010)	(0.013)	(0.013)	(0.013)
Parental strictness	-0.090	-0.084	-0.087	-0.014	-0.044	-0.042	-0.044
	(0.008)	(0.008)	(0.008)	(0.005)	(0.006)	(0.006)	(0.006)
Parental support	0.047	0.040	0.037	0.018	0.015	0.019	0.016
	(0.008)	(0.008)	(0.008)	(0.006)	(0.006)	(0.006)	(0.007)
Harsh parenting	-0.047	-0.053	-0.049	-0.010	-0.024	-0.031	-0.030
	(0.016)	(0.016)	(0.016)	(0.009)	(0.012)	(0.012)	(0.012)
Parent uni. aspirations	0.341	0.257	0.249	-0.010	0.121	0.096	0.094
	(0.017)	(0.017)	(0.017)	(0.012)	(0.013)	(0.013)	(0.013)
School environment	-0.001	-0.008	-0.016	0.021	0.018	0.018	0.012
	(0.008)	(0.008)	(0.008)	(0.005)	(0.006)	(0.006)	(0.006)
Classroom hard to manage	-0.061	-0.054	-0.044	0.006	-0.036	-0.032	-0.012
	(0.021)	(0.018)	(0.021)	(0.014)	(0.016)	(0.016)	(0.015)
Teacher engagement	0.037	0.039	0.041	0.015	0.012	0.015	0.014
	(0.008)	(0.008)	(0.008)	(0.005)	(0.006)	(0.006)	(0.006)
Teacher tired of teaching	0.016	0.021	0.031	-0.007	0.002	0.005	0.021
	(0.020)	(0.017)	(0.018)	(0.014)	(0.015)	(0.015)	(0.015)
Wave 1 inputs	No	Yes	Yes	FD	No	Yes	Yes
Fixed effects	None	None	School	Child	None	None	School
Schools	231	231	231	231	231	231	231
Classrooms	833	833	833	833	833	833	833
Students	11,581	10,831	10,831	10,831	11,581	10,831	10,831
Adjusted R^2	0.417	0.456	0.471	0.019	0.661	0.665	**0.676**
RMSE	0.757	0.731	0.721	0.648	0.578	0.574	**0.564**
CV-RMSE	0.760	0.729	0.769	0.984	0.576	**0.572**	0.574

This table shows coefficient estimates of regressing student test scores in wave 2 on educational inputs in wave 2 and wave 1 controls that differ according to the model in our trimmed sample with complete information on covariates of 231 schools and 833 classrooms and 11,581 students. Each column corresponds to a different value-added specifications following the categorization by Todd and Wolpin (2007) Each row present coefficients of educational inputs. The child fixed effects specification presents a first-difference (FD) estimator between waves 2 and 1. All models control all 17 pre-assignment characteristics tested for balancing and the leave-one-out mean of peer characteristics. Missing covariates and wave 1 inputs are imputed at the median and a missing covariate/input flag is always added. The last three rows report goodness of fit statistics for the different value-added models. The out-of-sample cross-validated RMSE (CV-RMSE) is calculated for each model using 6-fold cross validation, averaged over 5 random data partitions at the school level. The best fitting model according to each criterion is highlighted in ***bold****.*

Figure B.2: The Effect of Peer Test Scores in Wave 1 on Students' Own Test Scores in Wave 2 Mediated by Effects on Educational Inputs in Wave 2



This figure reports the mediated effects based on the decomposition of our academic peer effect estimate using only within-school variation in our trimmed sample with complete information on covariates of 231 schools and 833 classrooms and 11,581 students. These estimates are produced using a modified version of the Stata's `b1x2` *package by Gelbach (2016). Rows present the mediated effect of different educational inputs in wave 2. The ability peer effect estimate in this mediation exercise is shown in dark blue and differs slightly from our main estimate because we exclude a few observations with missing data on wave 2 educational inputs, and because we additionally control for wave 1 educational inputs. The total mediated effect is shown in green, and student, parent, and school & teacher inputs are shown in navy blue, maroon, and teal. All models control for student's own test scores in wave 1, all 17 pre-assignment characteristics tested for balancing, the leave-one-out mean of peer characteristics, and school fixed effects. All models also control for educational inputs in wave 1, which means we implicitly use a cumulative value-added model with school fixed effects to produce mediation estimates. Standard errors are clustered at the classroom level, and estimates different from zero at the 99, 95 and 90 percent confidence level are marked with ***, **, and *.*

Table B.16: The Effect of Peer Test Scores in Wave 1 on the Estimated Return to Educational Inputs in a Cumulative Value-Added model with School Fixed Effects

Outcome:	*Student test scores in wave 2 [std]*	
	Coef.	*Std. err.*
Value-added coef. interaction of peer test scores [std] with:		
Study time	-0.013*	(0.007)
Truant behavior	-0.023*	(0.013)
Cheated on exams	0.006	(0.012)
Academic self-efficacy	0.001	(0.007)
Mental health	-0.005	(0.007)
University aspirations	0.022	(0.015)
University expectations	0.024	(0.015)
Private tutoring	-0.020***	(0.007)
Time with parents	-0.009	(0.006)
Conflict with parents	0.006	(0.013)
Parental strictness	-0.009	(0.006)
Parental support	-0.012*	(0.007)
Harsh parenting	-0.010	(0.013)
Parent uni. aspirations	0.019	(0.014)
School environment	0.009	(0.006)
Classroom hard to manage	-0.059***	(0.017)
Teacher engagement	0.006	(0.007)
Teacher tired of teaching	-0.011	(0.016)
F-test interactions (p-value)	<0.001	
Wave 1 inputs	Yes	
Fixed effects	School	
R^2	0.69	
Schools	231	
Classes	833	
Students	11,581	

*This table shows interaction coefficient estimates of regressing student test scores in wave 2 on educational inputs in wave 2, all interacted with peer test scores in wave 1, in our trimmed sample with complete information on covariates of 231 schools and 833 classrooms and 11,581 students. Rows present coefficients of different educational inputs in wave 2 interacted with peer test scores in wave 1. All models control for student's own test scores in wave 1, all 17 pre-assignment characteristics tested for balancing, the leave-one-out mean of peer characteristics, wave 1 educational inputs and school fixed effects. Missing covariates and wave 1 inputs are imputed at the median and a missing covariate/input flag is always added. Standard errors are clustered at the classroom level, and coefficients statistically different from zero at the 99, 95 and 90 percent level are marked with ***, **, and *.*

Table B.17: 35 Education Systems in Comparative Perspective in TIMSS 1999

	Average class size	Student-to-teacher ratio	School-days yearly	Study hours daily	Percentage of: Math in groups	absent daily	school dropout	parents monitoring	5+ y. exp. teachers	weekly+ class disr.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Australasia and Pacific										
Taiwan	39	18	221	2.0	12.5	1.3	1.7	33.4	67.7	30.1
Japan	36	20	223	1.7	12.5	3.0	0.3	4.0	32.3	4.5
South Korea	42	24	225	1.6	18.5	0.5	0.8	9.2	20.6	43.0
Hong Kong	39	20	176	1.6	6.9	1.4	1.9	6.9	64.7	36.2
Singapore	37	20	180	3.5	32.5	1.6	1.0	1.8	49.1	32.2
Indonesia	42	23	251	3.0	27.5	2.7	1.9	74.6	75.6	20.5
Malaysia	38	19	198	3.8	38.2	3.8	3.9	28.5	42.4	25.6
Philippines	51	35	204	3.3	53.0	5.1	8.5	28.0	65.0	27.4
Thailand	39	31	202	2.9	51.1	3.7	3.1	57.0	64.7	12.6
New Zealand	26	16	188	2.0	43.5	6.9	7.9	0.8	55.7	68.1
Australia	26	16	196	2.0	27.8	7.6	6.0	0.1	57.7	73.1
Europe and Central Asia										
Netherlands	25	17	191	2.2	27.8	3.0	2.0	1.0	77.5	76.1
Belgium	19	10	175	3.0	7.3	3.0	2.2	3.6	74.1	40.4
Italy	20	13	210	3.6	27.7	6.3	1.5	20.7	63.7	46.6
Cyprus	29	14	160	2.8	16.8	3.1	2.2	4.3	23.1	54.5
Finland	20	12	186	1.8	18.5	5.2	1.4	1.5	70.7	50.0
Latvia	22	13	176	3.0	11.3	5.2	2.0	17.1	71.4	37.5
Moldova	26	17	205	3.3	31.2	4.5	2.7	34.8	78.0	29.1
Czech	24	19	197	1.9	8.5	7.7	1.8	4.8	66.1	62.9
Hungary	32	12	185	2.8	7.3	4.7	2.2	3.8	71.6	41.2
Bulgaria	21	15	172	3.0	27.9	6.9	2.5	8.6	71.2	22.0
Romania	24	19	159	3.4	23.9	3.5	1.9	12.3	65.9	17.0
Slovak	25	18	194	2.3	15.9	7.0	1.7	3.1	66.8	59.8
Slovenia	22	14	175	2.5	10.3	3.0	1.2	12.0	75.3	61.1
Macedonia	27	21	176	3.5	41.8	1.9	1.5	15.5	78.7	13.1
Russia	24	15	195	3.2	15.6	4.2	2.5	32.8	73.6	13.4
Turkey	39	63	181	3.6	22.2	3.2	4.9	32.2	52.0	14.5
North America										
Canada	27	20	188	2.2	40.1	5.4	5.5	4.6	58.2	60.3
United States	26	18	180	2.1	44.9	5.6	9.0	7.7	60.9	69.3
Central and Latin America										
Chile	34	37	193	2.4	57.6	6.4	5.1	2.5	67.1	45.9
Middle East and Africa										
Iran	32	27	209	4.1	38.6	2.4	2.3	27.5	26.7	21.4
Israel	34	14	199	2.7	39.8	5.1	1.6	3.5	61.5	60.8
Jordan	35	23	191	3.8	50.6	2.9	3.4	21.3	55.2	27.6
Morocco	28	24	207	3.3	44.5	4.0	7.6	11.0	70.7	31.5
South Africa	48	37	194	3.1	53.3	8.3	8.3	40.0	68.0	38.6
Tunisia	34	23	205	3.7	24.3	2.4	2.3	61.1	26.2	54.0

This table compares key features of Taiwanese junior high schools with those in 34 other countries participating in TIMSS 1999. The TIMSS 1999 data is publicly available through the TIMSS 1999 International Database. This table presents means using sampling weights and Jackknife repeated replications, following the TIMSS 1999 User Guide. All features presented here are reported by school principals, except daily study hours and Math taught in small groups, which is reported by students.

Appendix C 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; Run preliminary PCA on remaining variables
2. Count number of missing values by individual across variables
3. Standardize each variable, construct preliminary index as row-mean across standardized variables
4. Cut preliminary index into deciles: construct bins of similar input
5. For each variable, construct median within index decile among people used for imputation. If item is missing less than 1/3 of observations, replace missing value by median within index decile.
6. Re-run PCA now using variables with imputed values, to check visually that factor with and without imputed values have same distribution.

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

- the items used, and the corresponding respondent (Teacher, Student or Parents),
- the initial number of observations for each of these items separately,
- PCA factor loadings before and after imputation,
- the number of observations for the factor before and after imputation,
- the eigenvalue of the first and second factors before and 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 C.1: Construction of Standardized Scales of Educational Inputs in TEPS

Scale and Survey items used in scale	Obs. (1)	Resp. (2)	Factor loadings: Original (3)	Factor loadings: Imputed (4)
Study time				
Weekly hours at school on a normal school week	18,816	S	0.16	0.16
Weekly hours at school tutoring on a normal school week	18,759	S	0.28	0.28
Weekly hours on internet for homework on a normal week	18,703	S	0.10	0.11
I took the initiative to take tutoring out of school	18,679	S	0.35	0.35
Daily study hours, excl. school and tutoring time	18,701	S	0.54	0.53
I always make a study plan and review lessons according to it	18,618	S	0.57	0.57
I always take notes & make outlines while studying to review	18,626	S	0.62	0.62
I reduce leisure activities when preparing for exams	18,667	S	0.54	0.55
I take initiative to participate in academic competitions	18,544	S	0.29	0.30
Number books borrowed from school library since Junior High Grade 8	18,615	S	0.19	0.19
I did academic activities (incl. academic summer camp) over past summer	18,806	S	0.34	0.35
Factor observations			17,741	18,833
First factor eigenvalue			1.76	1.78
Second factor eigenvalue			0.44	0.42
Self-efficacy				
I am good at presentations or expressing my points of view	18,686	S	0.54	0.53
I am good at coordinating with other people in a group	18,744	S	0.58	0.58
I can plan things well no matter how trivial they are	18,731	S	0.65	0.64
I cooperate with everyone very well	18,709	S	0.55	0.54
I always come up with solutions to problems	18,708	S	0.62	0.62
My friends think of me as a person who always has lots of ideas	18,606	S	0.54	0.54
Factor observations			18,384	18,795
First factor eigenvalue			2.02	2.01
Second factor eigenvalue			0.05	0.05
Mental health				
How often feeling down or frustrated	18,716	S	0.71	0.71
How often want to scream or smash something	18,712	S	0.67	0.67
How often feeling body shaking, unable to focus	18,695	S	0.62	0.62
How often feeling lonely	18,676	S	0.64	0.64
How often feeling that you have bad fortune	18,658	S	0.59	0.59
How often feeling easily irritated by others	18,682	S	0.62	0.62
How often guilty, regret over some things	18,654	S	0.58	0.58
Factor observations			18,355	18,782
First factor eigenvalue			2.82	2.83
Second factor eigenvalue			0.24	0.24

This table presents detailed factor loadings and the number of observations used, before and after imputation, in the construction of our summative scales. Col. (1) reports the initial number of complete observations available, Col. (2) indicates whether teachers (T), parents (P) or students (S) respond to each item. Cols. (3) and (4) report factor loadings on the first factor, before and after imputation. See Appendix C for details on our imputation procedure.

Table C.1: Construction of standardized scales of educational inputs in the TEPS data (continued)

Scale and Survey items used in scale	Obs. (1)	Resp. (2)	Factor loadings: Original (3)	Factor loadings: Imputed (4)
Private Tutoring				
Hours per week spent on tutoring outside school	18,747	P	0.78	0.78
Monthly expenditures this semester for this child's tutoring	18,755	P	0.78	0.78
Factor observations			18,586	18,916
First factor eigenvalue			1.21	1.22
Second factor eigenvalue			-0.21	-0.20
Time with Parents				
Weekly number of dinners with the child	18,783	P	0.44	0.45
Spouse: Weekly number of dinners with the child	18,493	P	0.44	0.45
Factor observations			18,457	18,819
First factor eigenvalue			0.39	0.41
Second factor eigenvalue			-0.21	-0.21
Parent strictness				
How many of your parents set strict rules for your daily routine?	18,828	S	0.61	0.61
How many of your parents set strict rules about spending money?	18,819	S	0.54	0.54
How many of your parents set strict rules about demeanor?	18,806	S	0.63	0.63
How many of your parents set strict rules about health habits?	18,731	S	0.60	0.60
How many of your parents set strict rules about making friends?	18,821	S	0.57	0.57
How many of your parents uses guilt and emotional blackmail?	18,821	S	0.51	0.51
How many of your parents does not allow you to argue with them?	18,816	S	0.50	0.50
How many of your parents discipline you very strictly?	18,809	S	0.53	0.53
Factor observations			18,648	18,831
First factor eigenvalue			2.54	2.55
Second factor eigenvalue			0.15	0.15
Parent emotional support				
My parents pay attention to my ideas and thoughts	18,816	S	0.66	0.66
I seek my parents' help when I encounter difficulties	18,811	S	0.67	0.67
My parents accept me as I am	18,799	S	0.62	0.62
Factor observations			18,769	18,827
First factor eigenvalue			1.27	1.27
Second factor eigenvalue			-0.15	-0.15

Note: This table presents detailed factor loadings and number of observations used in the construction of our summative scales as imputation procedure. Col. (1) reports the initial number of complete observations available, Col. (2) indicates whether teachers (T), parents (P) or students (S) respond to each item, Cols. (3) and (4) report factor loadings on the first factor respectively before and after imputation. See Appendix C for details about our imputation procedure.

Table C.1: Construction of standardized scales of educational inputs in the TEPS data (continued)

Scale and Survey items used in scale	Obs. (1)	Resp. (2)	Factor loadings: Original (3)	Factor loadings: Imputed (4)
School Environment				
My school's requirements on students are quite reasonable	18,614	S	0.39	0.39
My school is fair in terms of rewards and grading	18,741	S	0.46	0.46
The campus of my school is safe	18,709	S	0.56	0.56
My school cares about their students	18,340	S	0.62	0.62
My school has a great atmosphere for learning	18,690	S	0.52	0.52
Factor observations			18,053	18,814
First factor eigenvalue			1.33	1.34
Second factor eigenvalue			0.015	0.013
Teacher Engagement				
How many teachers talk about people skills in class	18,795	S	0.70	0.70
How many teachers often discuss life goals, do career advice	18,784	S	0.73	0.73
How many teachers often recommend books, encourage reading	18,783	S	0.62	0.62
How many teachers often use real life and practical examples	18,772	S	0.62	0.62
How many teachers take free time to help students with personal issues	18,795	S	0.53	0.53
How many teachers often use guilt or emotional blackmail	18,784	S	0.45	0.45
How many teachers praise me when I study hard	18,744	S	0.53	0.53
Factor observations			18,590	18,820
First factor eigenvalue			2.56	2.56
Second factor eigenvalue			0.17	0.17

Note: This table presents detailed factor loadings and number of observations used in the construction of our summative scales as imputation procedure. Col. (1) reports the initial number of complete observations available, Col. (2) indicates whether teachers (T), parents (P) or students (S) respond to each item, Cols. (3) and (4) report factor loadings on the first factor respectively before and after imputation. See Appendix C for details about our imputation procedure.

Appendix D An Application to Ability Peer Effects: Sensitivity Analyses and Additional Results

D.1 Sensitivity Analyses Related to our Identification Strategy

D.1.1 Permutation-Based Balancing 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 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); Lim and Meer (2017, 2020) in our trimmed sample.

For these tests, we simulate 1,000 classrooms under the null of random assignment of students to classrooms within schools. We do so by drawing students from schools with replacement and randomly assigning them to classrooms while keeping the core structure of the data (i.e., respecting students' assignment to schools, and number and size of classrooms within each school). We then calculate the mean of our 18 pre-assignment characteristics in each of the 1,000 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. 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 B.5 shows these permutation-based empirical p-values for each pre-assignment 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 implementation of random assignment to classrooms within schools in our trimmed sample.

D.1.2 Non-Parametric Balancing Tests

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 (e.g., Ammermueller and Pischke, 2009; Sojourner, 2013; Feld and Zölitz, 2017).

We implement this test in the trimmed TEPS data 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 B.6 shows empirical p-value distributions for each characteristic separately. Consistent with our other tests, these results also show evidence of minor imbalances on some characteristics. Overall, however, these tests provide again evidence in support of random assignment to classrooms within schools in our trimmed sample.

D.1.3 Different thresholds for the Fishing algorithm

An important and somewhat arbitrary decision in implementing our Fishing Algorithm is deciding when to classify any particular school as not compliant with the mandate of random assignment. Recall from Section 2.2 that we do so based on $\hat{P}_{sl}$, the posterior probability that school s belongs to the latent class of sorter schools—which we will refer to as the latent sorter probability, for short. Our intuitive thumb rule is: a school is a sorter if its latent sorter probability is larger than all other latent class probabilities combined. However, this is not the only way to classify such schools. Another approach is to pick a fixed probability threshold and consider any school with a latent sorter probability above that threshold as a sorter.

In Appendix Table B.7 we show how all our main results on achievement peer effects change had we implemented this fixed threshold approach at different levels, ranging from 0.2 (relatively strict, removing schools that are even somewhat likely be sorters) to 1 (very relaxed, effectively removing only schools for which S_s is exactly equal to 1). The bottom of the table shows that the sorting statistic of Jochmans (2023) grows monotonically with the threshold, as expected, and starts rejecting the null of no sorting for thresholds of 0.8 and above. For thresholds below 0.8, we find consistent achievement peer effects on test scores of between 3.5 and 4.3 percent of a standard deviation, as well as consistent positive effects on time with parents and harsh parenting, and negative effects on school environment and teachers' reports of difficulties with classroom management. In general, for thresholds 0.8 and below coefficients are very stable and compare well to our main effects. For a threshold of 1, however, we find larger achievement peer effects on test scores, smaller effects on time with parents, harsh parenting and school environment, similar effects on difficulties with classroom management, and a new positive effect on parental support. Most of these differences are consistent with stronger ability sorting into classrooms for these schools. By and large, however, Appendix Table B.7 shows that our results are not overly sensitive to which threshold we use in the Fishing Algorithm as long as the resulting estimation sample passes Jochman's sorting test.

D.1.4 Results weighted by the probability of being a sorter school

One could view the Fishing Algorithm as a way to solve a trade-off between bias and variance. Stricter school selection rules in the form of lower thresholds τ—the probability threshold under which a school is considered a sorter—favor unbiasedness (since they are more likely to remove sorter schools) at the

expense of statistical power (since less data are used to produce estimates). Relaxing selection rules, in the form of higher τ, produces more precise estimates at the expense of a higher chance of using data from a sorter school as part of the estimation sample. A different approach to this trade-off is not set a rule based on τ but rather to use all the available data and weight each observation by $1 - \hat{P}_{sl}$. This approach effectively uses data from all schools except those for which $\hat{P}_{sl}$ is equal to one—the most egregious sorter schools as detected by the algorithm—to estimate peer effects, but gives lower weight to schools that are detected as more likely to be sorters.

Appendix Table B.8 shows sorting and balancing tests using the complete TEPS data but weighting each observation by $1 - \hat{P}_{sl}$. The results in this table are based on up to 19,957 student observations, a 56 percent increase in the sample sized when compared to the 12,793 observations used in Table 3. Using these weights we see no evidence of sorting on test scores and little evidence of failures in balancing. We do see that this weighting does not eliminate imbalances in whether parents work in government and whether students report being enrolled in a gifted academic classroom, and it is still unsuccessful in eliminating imbalances on whether parents attempted to get their children into a better classroom. Yet overtrimming issues are less severe using this method (as evidenced by comparing e.g., the coefficient on family income in this table with that of tables 3 and B.8). More importantly, using these weights in the complete TEPS data meaningfully increases statistical power, reducing standard errors in Panel B by around 21 percent on average. All these results are an excellent demonstration of the different way in which weighting solves the bias-variance trade-off compared to the main approach in the Fishing Algorithm.

Appendix Table B.9 shows our main results using the weighted complete TEPS data. We see evidence of achievement peer effects of 0.041 SD, statistically indistinguishable from the 0.037 SD effects in our preferred specification. We also see very similar effects of higher achieving peers on time spent with parents, on harsh parenting, on classrooms being hard to manage, and on the quality of the school environment (with the former three also being statistically significant at standard levels). We also see no evidence here of peer effect on any other educational input. Importantly, the standard errors of these estimates are, on average, around 17 percent smaller than those from Equation 3 and behind Figure 4. Overall, using $\hat{P}_{sl}$ as estimated data weights seems to be a promising alternative use of our Fishing Algorithm, bringing meaningful increases in statistical power without introducing any evident bias in our results.

D.2 Robustness to Measurement Error and Classroom Sampling

D.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.[21]

The leftmost two columns of Appendix Table B.10 shows that the peer effects we identify are mostly driven by the mathematical subcomponent of the comprehensive ability test scores. The rightmost three columns show that using the Bayesian posterior means of the IRT models as measures of achievement leads to statistically indistinguishable effect sizes–perhaps a bit smaller—and that there are no efficiency gains from using these measures over the simpler test scores.

D.2.2 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 classroom 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 rest of the columns implement specifications which trade off more power for less bias reduction, from left to right.

Appendix Table B.11 shows substantially larger effects of higher-achieving peers on student test scores. The largest increase in academic peer effect estimates is produce by the recommended controls for the share of classroom peers observed, at 11.1 percent of a standard deviation. This is an almost fourfold increase from our preferred specification, which points to a substantial incomplete sampling downward bias in our main estimates. The similarities with a standard attenuation bias are consistent with the random assignment of students to classrooms within schools in our trimmed sample. The less restrictive estimators produce slightly more modest point estimates of around 8 percent of a standard deviation, remarkably consistent across specifications. Nevertheless, these point estimates remain within the range of estimates found in previous studies, especially considering that peers here have had two years to work their effect on student achievement.

The Sojourner corrections mostly result in result similar but less precisely estimated effects of higher ability peers on educational inputs. The one exception is the negative effects on teachers' reports that classrooms are hard to manage. For this outcome, the corrections also result in a threefold increase in the effect size, again consistent with an attenuation-like bias from incomplete peer sampling. The analyses do not reveal other effects of higher-ability peers. There is no evidence of effects on any other educational input once estimates have been corrected.

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

Together, these estimates suggest that incomplete peer sampling is severely biasing our per effects downwards, and suggests that the one empirically meaningful mechanisms for these effects is the ease of classroom management.[22]

D.3 Randomization Inference and Multiple Hypothesis Testing

Having established the robustness of our point estimates, we now reassess our statistical inference. We first do so using the randomization-*t* procedure from Young (2019). 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 correcting for few treatment clusters or issues of joint testing are less important for this study, since we observe several classrooms per school and each regression has only 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 and our Fishing Algorithm, where we keep school structure intact but simulate random assignment of students to classroom within schools. The key difference here is that, in each simulation, we capture the *t*-statistics of interest—the coefficient of average peer achievement divided by its cluster-robust standard error—and construct empirical *p*-values based the share of instances 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. The leftmost column of Appendix Table B.12 shows that when using randomization-*t* *p*-values for conducting inference, we still find statistically significant effects of higher-achieving peers on student achievement and on difficulties with classroom management, but no longer find statistically significant effects on any other educational input.

In a second analysis, we adjust our inference to control for the Family-wise Error Rate (FWER): the probability of incorrectly rejecting at least one null hypothesis, given that all the null hypotheses are true. We define two "families" of tests for this correction; one for asking whether higher-achieving peers affect student test scores in the TEPS data, and another for asking whether the effect of higher-achieving peers on test scores operates through the educational inputs measured in the TEPS data. Following Rubin (2021), we note that the first is an individual test and therefore the FWER correction should not be applied. We could define the second test as a disjunction test (where we reject the joint null hypothesis if *any* of the constituent tests—read: effects of higher-achieving peers on educational inputs—is rejected) or a conjunction test (where we reject the joint null if all the constituent tests are rejected). The exploratory nature of our analysis of effects on educational inputs leads us to opt for a disjunction test. We therefore implement the FWER control on all our effects on educational inputs via the Romano-Wolf multiple hypothesis correction (Romano and Wolf, 2005*a*,*b*) using the `rwolf` Stata command (Clarke, Romano and Wolf, 2019). The rightmost columns of Appendix Table B.12 shows that our disjuntion test cannot

22. There are other potential issues with incomplete classroom sampling, especially if our peer effects varied with classroom or school size, if the classroom sampling rate were correlated with our regressors, or if our Fishing Algorithm were selecting schools with different average sampling rates. Fortunately, none of these occur in our data.

reject the null hypothesis that none of the educational inputs measured in TEPS can help explain the effect of higher-achieving peers on student's test scores.

Overall, with these different inference methods we still find strong evidence of achievement peer effects in our data but much weaker evidence of significant effects on educational inputs.

D.4 Additional Results

In this section we first explore the potential heterogeneity in our achievement peer effect estimates. Previous studies have found heterogeneity in achievement peer effects across many dimensions, including ability (Carrell, Fullerton and West, 2009), gender (Whitmore, 2005; Lavy and Schlosser, 2011) and race (Hoxby, 2000; Hoxby and Weingarth, 2005).

D.4.1 Heterogeneity of Effects

There are countless dimensions to explore heterogeneity in achievement 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 explore heterogeneity in the effect of higher-achieving peers on test scores across student and peer ability, student gender, household income, parental education, and teacher experience. Appendix Table B.13 shows that, by and large, there is little subgroup heterogeneity in our estimated academic peer effects, and a joint test cannot reject the null of no heterogeneity across all these dimensions (p-value = 0.877). Effects seem larger for students assigned to less high-achieving peers, which is consistent with decreasing marginal returns in peer achievement. Effects also seem slightly larger for more disadvantaged students in terms of household income and parental education. However, differences across these subgroups are imprecisely estimated.[23]

D.4.2 Effects in the Long-Run

Next, we explore the effects that higher-achieving peers can have on students' long-term outcomes. We first estimate effects on student standardized test scores in the first and third year of high school using data from waves 3 and 4 of the TEPS junior high school cohort. Unfortunately, waves 3 and 4 only follow a smaller subset of around 3,000 students from the original cohort, and the followed students are not random. Female students, students with more educated parents, and higher-scoring students are more likely to be followed into waves 3 and 4. Fortunately, this student attrition is not related to peer achievement (p-value = 0.482). We then estimate effects on educational attainment and employment around ages 20 and 24 using data from the TEPS Beyond project. The TEPS Beyond project follows roughly the same students in waves 3 and 4 of the TEPS after graduation, and attrition into this sample is also unrelated to peer achievement (p-value = 0.485). Table B.14 shows either zero or negative effects of higher-ability peers on test scores during high school. However, these effects are very imprecisely estimated; their ex-post MDE

23. We also see little heterogeneity in the effects of higher-achieving peers on educational inputs. The only statistically significant heterogeneity at the 95 percent level is on effects on parents' aspirations for their children to go to university (p-value = 0.032). These results show more positive effects on parents' aspirations for boys, higher income students, and in classrooms with less experienced Dao Shi.

is larger than 10 percent of a standard deviation. This means that, unfortunately, we cannot say much about longer-term effects of higher-ability peers on achievement. On educational attainment, however, we do see that students exposed to higher-ability peers are more likely to be attending or finished university and less likely to be attending or finished vocational education around age 20. This hints at positive effects of higher-ability peers on academic attainment in line with their effects on test scores in junior high school. Finally, we see no evidence of longer run effects of higher-achieving peers on postgraduate educational attainment, employment or income.

D.4.3 Mediation Analysis

Lastly, we formally calculate the extent to which the effects of higher-achieving peers on test scores are explained by their effect on educational inputs. We note that these results can only be interpreted as capturing causal mediating effects under strong, perhaps heroic assumptions. To perform this analysis, we follow the decomposition in Gelbach (2016), which we adapt 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 as:

$$ME = \sum_k ME_k = \sum_k \underbrace{\frac{\partial Ed.Inputs^k_{ics2}}{\partial \overline{TestScores}^{-i}_{ics1}}}_{\text{(A)}} \times \underbrace{\frac{\partial TestScore_{ics2}}{\partial Ed.Inputs^k_{ics2}}}_{\text{(B)}}$$

where 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 and the term (B) 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) (see e.g., Todd and Wolpin, 2003, 2007), yet in Table B.15 we explore most feasible value-added alternatives in our data and show that they yield largely similar return estimates. Based on goodness of fit criteria and in line with our setting, our preferred specification uses a cumulative value-added model with school fixed effects. Figure B.2 shows that using this model our estimated ME is actually slightly negative. This conclusion is robust to using any of the other value-added estimated returns in Table B.15.

Finally, we consider the possibility that higher-achieving peers improve student test scores by improving the technology of skill formation rather than by changing educational inputs. We approach this by estimating whether there is heterogeneity of the value-added estimates of the return to educational inputs across the peer ability distribution. Such heterogeneity would match existing evidence that e.g., higher-achieving classrooms change the productivity of some teaching practices (Aucejo et al., 2020). Table B.16 shows that exposure to higher-achieving peers does have an impact on value-added estimates of the returns to educational inputs (p-value < 0.001). Higher-achieving peers lower the return to private tutoring investments by two percent of a standard deviation (a 51 percent decrease from the average estimate in the rightmost column of Appendix Table B.15) and increases the harm done by hard-to-manage classrooms by 5.9 percent of a standard deviation (from effectively zero). Combined with the effects on educational inputs shown in Figure 4, however, it is unclear whether these changes in technology make it easier or harder to explain our peer effects on student test scores.

Overall, subgroup heterogeneity in peer effects on test scores, educational inputs of the technology of skill formation is not a likely to be an explanation for the fact that our many educational inputs do not mediate academic peer effects.