
Working Paper No. 189

# The Earth is Not Flat: A New World of High-Dimensional Peer Effects

Aurélien Sallin and Simone Balestra

Universität Zürich
IBW – Institut für Betriebswirtschaftslehre

UNIVERSITÄT BERN

Working Paper No. 189

# The Earth is Not Flat: A New World of High-Dimensional Peer Effects

Aurélien Sallin and Simone Balestra

January 2022

Please cite as:
"The Earth is Not Flat: A New World of High-Dimensional Peer Effects " Swiss Leading House "Economics of Education" Working Paper No. 189, 2022. By Aurélien Sallin and Simone Balestra.



---

The Swiss Leading House on Economics of Education, Firm Behavior and Training Policies is a Research Program of the Swiss State Secretariat for Education, Research, and Innovation (SERI).

www.economics-of-education.ch

# The Earth is Not Flat:

# A New World of High-Dimensional Peer Effects *

Aurélien Sallin† and Simone Balestra‡

**Abstract**

The majority of recent peer-effect studies in education have focused on the effect of one particular type of peers on classmates. This view fails to take into account the reality that peer effects are heterogeneous for students with different characteristics, and that there are at least as many peer effect functions as there are types of peers. In this paper, we develop a general empirical framework that accounts for systematic interactions between peer types and nonlinearities of peer effects. We use machine-learning methods to (i) understand which dimensions of peer characteristics are the most predictive of academic success, (ii) estimate high-dimensional peer effects functions, and (iii) investigate performance-improving classroom allocation through policy-relevant simulations. First, we find that students' own characteristics are the most predictive of academic success, and that the most predictive peer effects are generated by students with special needs, low-achieving students, and male students. Second, we show that peer effects traditionally reported by the literature likely miss important nonlinearities in the distribution of peer proportions. Third, we determine that classroom compositions that are the most balanced in students' characteristics are the best ways to reach maximal aggregated school performance.

**Keywords**: peer effects, high dimensionality, machine learning, classroom composition.

**JEL Classification**: C31, H75, I21, I28.

*We are grateful to Beatrix Eugster, Petra Thiemann, participants of the Brown Bag seminar at Lund University for their helpful comments. All code is available on the repository `https://github.com/ASallin/hdpx`. This study was conceived and drafted when Dr. Balestra was employed at the University of St. Gallen, and the findings and views in this manuscript do not necessarily reflect the official views or policy of his current employer. The usual disclaimers apply.

†University of St. Gallen. Email: aurelien.sallin@unisg.ch

‡University of St. Gallen. Email: simone.balestra@gmail.com

# 1 Introduction

In the past 15 years, the literature on peer effects in the school environment has grown exponentially (e.g., Angrist, 2014; Balestra, Sallin, and Wolter, forthcoming; Bifulco, Fletcher, and Ross, 2011; Black, Devereaux, and Salvanes, 2013; Brenøe and Zölitz, 2020; Burke and Sass, 2013; Carrell and Hoekstra, 2010; Carrell, Hoekstra, and Kuka, 2018; Lavy, Paserman, and Schlosser, 2012; Lavy and Schlosser, 2011), and we now have a better understanding of the main driving forces at play in the classroom. However, most of the classroom peer effects[1] have been studied "in isolation", i.e., researchers have focused on the effect of one particular peer characteristic on their classmates (for instance, the effect of female peers on their classmates). These analyses have two main limitations: they do not sufficiently account for the fact that spillovers have heterogeneous impacts on individuals with different characteristics, and they do not incorporate the reality that there are *at least* as many spillovers as there are types of students. As a result, such analyses are often unable to capture the granularity of peer effects, and miss the fact that individual effects are important (Isphording and Zölitz, 2020). As a consequence, and from an estimation perspective, most peer-effect analyses rely on average effects estimated with linear-in-means models, and are therefore likely to miss the complexity of spillovers.

In this paper, we provide a more comprehensive analysis of educational spillovers by looking at the systematic interaction of peers and their many characteristics. We start with the assumption that "true" spillover effects are nonlinear and high-dimensional (Sacerdote, 2014). They are nonlinear as they vary with group composition. For instance, we do not expect the marginal effect of female peers to be similar when the proportion of female peers in the classroom is high or low. Moreover, the marginal effect of female peers is likely to affect low-achieving male students differently than high-achieving female students. In addition, spillover effects are high-dimensional: they interact with students' characteristics and with each other. For instance, the effects from female peers likely interact with the effects from younger peers, are different if they are generated by high- or low-ability female students, and impact high- and low-ability students differently. Indeed, the complexity of peer effect models increases with the number of students' characteristics that are thought to generate spillovers.

Using up-to-date machine learning (ML) methods, we develop a general empirical approach that systematically considers the nonlinearities and high-dimensionality of spillover functions. We determine which dimensions in the classroom are the most predictive of academic success, we develop a framework

[1] In the following of this paper, we use "peer effects", "spillovers" and "spillover effects" interchangeably.

that allows us to estimate high-dimensional spillover functions, and we conduct classroom allocation simulations (counterfactual analysis). There has been a surge in the use of highly predictive ML algorithms in economics and causal econometrics (e.g., Athey, 2019; Athey and Imbens, 2019; Mullainathan and Spiess, 2017), but no study so far has used such methods to address the problem of educational spillover effects. For this study, we use unique data on the universe of students in middle schools from the canton of St. Gallen, the fifth-largest state in Switzerland. Merging student achievement data from a mandatory standardized test in grade eight, and administrative records of the School Psychological Service (SPS), we are able to observe a rich set of students' characteristics. We focus on gender, age at test (which is a good proxy for achievement, see Bietenbeck, 2020), whether the student has special needs (see Balestra, Eugster, and Liebert, forthcoming), whether the student is a high-ability (gifted) student (see Balestra, Sallin, and Wolter, forthcoming), and finally whether the student is a nonnative speaker. For identification, we exploit random variation in cohort composition within school-tracks, and in classroom composition within school-track-years. To test the random allocation of students to cohorts within school-tracks, and to classrooms within school-track-years, we adapt balancing checks traditionally used in the peer effect literature to a setting with more than one characteristic of interest. Following these randomization tests, we argue that both the distribution of students to cohorts within school-tracks and the random allocation of students to classrooms within school-track-years are consistent with random variation for all measured students' characteristics.

While the advantage of ML algorithms is not always clear in empirical studies, the problem at hand ideally calls for the use of such methods. One main advantage of ML methods is that they can handle high-dimensional models with a very large set of predictors by selecting predictors that are highly predictive of a given outcome. We conduct the following thought experiment: we imagine that school administrators observe a given set of student and cohort or classroom characteristics. Among all the characteristics they observe, they would like to know which ones are the most predictive of academic success, and in which combination. We run stable selection procedures (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013) with ML regularization algorithms to uncover these variables. These procedures have been used in biology (for identifying the risk factors of coronary heart disease, molecular features, etc.), and this is the first time such algorithms are used to help us understand social phenomena in a social science setting. We find that students' own characteristics are the most important predictors of their academic success, rather than students' peers: peer effects are rarely selected as predictors of academic performance, while selected variables are all students' own characteristics or interaction between differ-

ent own characteristics. The fact that interaction between characteristics are selected is a first indication that focusing on main characteristics (such as gender only) is likely to hide substantial heterogeneity. Qualitatively, we show that the few selected spillover effects are dominated by the effect from peers with special needs and from low-achieving peers. This confirms findings that older low-achieving peers and peers with special needs have an important impact on their peers (e.g., Balestra, Eugster, and Liebert, forthcoming; Bietenbeck, 2020). Finally, when we compare variables selected at the cohort and classroom levels, we observe that peer effects are more "diluted" at the cohort level, which is expected as peers exert most of their influence in smaller groups (this is consistent with Burke and Sass, 2013, who discuss how peer effects change between cohort and class level).

Another advantage of ML methods is their flexibility. Such data-driven algorithms are thus ideal to paint a more realistic picture of spillovers. We leverage cohort and classroom identification strategies to flexibly estimate high-dimensional peer effect functions. Moreover, we develop a procedure with ML algorithms that allows us to assess both peer effects at a granular level and heterogeneous spillover effects of particular finely defined groups of students on other groups of students. This flexible estimation procedure allows us to systematically recover results that correspond to, and confirm, main findings of the literature. However, our findings do uncover more interesting heterogeneities: at the cohort level, spillovers from gender status are virtually nonexistent. Similarly, we deliver a different picture of spillovers at the classroom level: effects generated from older students, students with SEN, and nonnative students, are downward sloping and not constant in the share of peers within classrooms. This means that the more older peers, peers with SEN, or nonnative peers in the classroom, the more negative their impact on other students and on themselves become. These findings suggest that classrooms environments that are less homogeneous in students' characteristics (i.e., classrooms that are more mixed and have low segregation along types of students) provide higher chances of academic success, and that even small variations in the proportion of these peers have large negative effects.

In a last step, we perform counterfactual analyses, in the spirit of Graham, Imbens, and Ridder (2010), Graham (2011), and Graham et al. (2020). We conduct simulations to investigate what would happen to a given student if he or she were put in a different classroom environment. We do so by computing counterfactual assignments under the constraint that the existing pool of students remains unchanged. For instance, we are interested in learning about aggregated outcomes when school administrators decide to slightly manipulate the classroom compositions by substituting, from each classroom, one student with a particular characteristic with another student, and to create "clusters" of students of a given type in

one classroom (case of "marginal segregation"). More precisely, we assess the average counterfactual effects (*ACE*) and the conditional average counterfactual effects (*CACE*) of these small manipulations of classroom composition, i.e., how much counterfactual allocations affect particular groups of students. The key takeaways from this counterfactual analysis are three. First, the results strongly suggest that marginally increasing segregation has different impacts depending on which students are segregated. Our results clearly show that creating clusters of students with SEN has large negative impact on the aggregated school performance. However, creating clusters of female students makes no difference at the aggregate level, as students greatly benefit from having more female students in their classroom while classrooms who lose female students perform less well. Second, students in classrooms containing clusters (thus classrooms that are more homogeneous in students' characteristics) are overly harmed by higher homogeneity in peers exerting negative peer effects, and the aggregated harm done to the segregated groups always exceeds the aggregated benefits for the mainstreamed students. This is because marginal spillover effects are not constant.[2] Finally, our results show – unsurprisingly – that marginal segregation increases overall inequality as measured by the Gini coefficient.

The contributions of the present study are four. First, this study is the first to use cutting-edge ML methods to examine peer effects. These flexible estimation methods allow us to investigate spillover effects at the most granular level, i.e., at the quasi-individual level of interacted types. To our knowledge, this is the first paper to attempt to estimate spillovers at such a granular level, perhaps with the exception of Isphording and Zölitz (2020) who measure the value-added of individual students placed in different peer groups.

Second, our findings confirm the shared intuition in the school peer-effect research that "not all peer effects are created equal." Using variable selection methods, we are able to offer a ranking of relevant peer characteristics, highlighting the fact that own characteristics remain the most important characteristics for academic achievement. We are also able to show that not all peer effects affect students with different characteristics in the same way.

Third, we show that spillover effects are nonlinear (high dimensional), supporting the evidence from the recent network literature (Bramoullé, Djebbari, and Fortin, 2020). In a way, we take the intuition of Sacerdote (2014) to the letter by systematically accounting for nonlinearities and high dimensionality

[2] In the appendix, we show that this the case by conducting the following thought experiment: we take clustering to the extreme and we set counterfactual classroom allocations in which we allow students to be completely segregated according to their types. We show that complete segregation is harmful to the segregated groups, and that this harm outweighs the benefits that the non-segregated students enjoy.

in our estimation procedure, instead of considering nonlinearities and high dimensionality as *ad hoc* phenomena, as it is usually done in the robustness sections of several peer-effect studies.

Finally, we discuss counterfactual allocations and show that, keeping resources fixed, marginally segregating students of a given type neither increases aggregate achievement nor fosters equality. This evidence adds to the policy debate on the value of school segregation, especially in the case of the segregation of students with SEN (see Balestra, Eugster, and Liebert, forthcoming; Sallin, 2021). Taken to their fullest, these last conclusions are supportive of inclusive education with respect to all observed types. We remain however cautious in devising policy recommendations, primarily because we perform our exercise in a world where resources are fixed and cannot be allocated – for example – to classes with more disadvantaged students. Nonetheless, we may deliver the following claim: simulations of counterfactual effects indicate that a composition of classrooms which is more homogeneous in students' characteristics does not improve the aggregate academic performance of students.

## 2 Background and Data

The education system in Switzerland offers ideal conditions to observe the diversity of students in general (mainstreamed) education. Almost all students in Switzerland complete compulsory education in public school; only 5% of students in a cohort go to private school. Importantly, students cannot freely choose their public school but are assigned to schools according to their municipality of residence. As inclusion is an important public school policy objective, children of different ability, gender, ethnicity, special needs, and socio-economic background are educated in an inclusive setting whenever possible.[3] This situation is comparable to the academic environment of most OECD countries, as well as the US.

The present study considers the universe of middle schools from the canton of St. Gallen, the fifth-largest state in Switzerland. St. Gallen follows the typical Swiss curriculum[4], which includes a two-year entry level (kindergarten) and nine years of compulsory schooling, the first six years of which are at the primary level and the last three at the middle school level. Crucially for our study, our setting stands out for the following three important reasons. First, classes remain unchanged within each level but

[3] Special schools in Switzerland represent 4.7% of all schools, serving 1.8% of the student population (Swiss Federal Statistical Office, 2020).

[4] Switzerland has a federal structure and gives the 26 cantons – regional administrative entities similar to U.S. States – some autonomy in educational policy decision-making. The degree of coordination among the cantons remains nonetheless relatively high: since the Intercantonal Agreement on the Harmonization of Compulsory Education in the 1970s, the cantons have applied virtually the same common curriculum for compulsory schooling.

are reshuffled when transitioning from primary to middle school. This means that, as opposed to the American setting and that of other countries, the class composition is stable during middle school and across school subjects. Second, tracking occurs at the end of primary school. There are two main tracks, the *Sekundarschule* (higher track) and the *Realschule* (lower track). Assignment to either track is based on student performance and the primary school teacher's recommendation. Third, regardless of the track, all students take a mandatory standardized test in core subjects at the end of grade eight – the second to last year of compulsory schooling. The test, named "Stellwerk," is computer-based and administered by the cantonal Department of Education. Stellwerk is a norm-referenced, self-scoring, adaptive exam similar in spirit to the Graduate Record Examination.[5]

Data for this study stem from two administrative sources. First, we use test score data for the population of students enrolled in eighth grade in the canton of St. Gallen during the years 2008 to 2017. These data are supplied by the Stellwerk test provider and contain information on student achievement (standardized test score in Math and German), student characteristics (date of birth, gender, and native German speaker), and composition of schools and classes in grade eight (school, track, classroom identifier, and teacher identifier). The second data source is the administrative records of the School Psychological Service (SPS) of the canton of St. Gallen. The SPS is a centralized service provider for all schools in the canton, divided into separate administrative units for the main city St. Gallen and the remainder of the canton, which is served by seven regional offices. The SPS provides diagnosis and counseling for children, parents, and teachers for school-related problems. In these data, we observe all students who were ever registered to the SPS, along with information on general counseling, diagnosis of learning disabilities, developmental deficiencies, conflict mediation, and schooling strategies for children with any kind of special educational needs.[6] From the SPS data, we can identify students with a special needs diagnosis (e.g., learning disability or socio-emotional problem) and students who are classified as intellectually gifted (i.e., IQ above 130 points).[7]

We merged the two administrative data sets using anonymous student identifier provided by the Swiss Statistical Office. We impose the following data restrictions. We remove 1,265 students with special needs assigned to special schools (fully segregated special education schools), and we remove 1,464 students who did not take the SW8 test. We additionally remove outliers in term of age at test (289

[5]For more details on Stellwerk, please refer to Balestra, Eugster, and Liebert (forthcoming) and Balestra, Sallin, and Wolter (forthcoming), from which we draw heavily in this section.

[6]Balestra, Eugster, and Liebert (2020) offer a comprehensive discussion of the SPS, its role in St. Gallen, and the detailed procedure of registration with the SPS.

[7]Balestra, Sallin, and Wolter (forthcoming) discuss in detail the procedure of giftedness assessment via quantitative (IQ) and qualitative testing at the SPS.

observations). We decide to remove all classrooms that have less than 9 students and classrooms that have more than 31 students (2,476 observations). Finally, to have enough within-school variation left, we decide to keep only schools that appear at least five times in the data (1,082 students).

[Insert Table 1 here]

Table 1 describes our dataset. The data contain 2,674 classrooms in 142 school-tracks over ten academic years. As individual binary characteristics, or "types", we focus on gender, nonnative status, relative age status, intellectual giftedness, and special needs diagnosis. Relative age indicates when a student is older than the typical age when taking the Stellwerk test. Older peers are students who have either repeated a grade or who have started school later. On average, 17% of students are older than 15 years old, i.e., the typical age at which students in St. Gallen take the test (for information, 25% of students are younger than the typical student when they take the test). We take the indicator of being older at test as a good proxy for low-achievement (Bietenbeck, 2020). Students identified with SEN are students who have been referred to the SPS in primary school and who received a diagnosis. We identify 29% of students with SEN, which is very high in international comparison. The reason for this high number is that we adopt a very broad definition of SEN: some of these students have severe SEN (such as physical handicaps), while other students are sent for milder SEN (e.g., counseling, tutoring, speech therapy, or learning disabilities). The gifted population is very small, and covers around 1% of the student population. These students are identified gifted students with an IQ score above 130 or with a qualitative assessment that confirms their intellectual giftedness. Fifteen percent of students are nonnative, meaning that they come from families who do not speak German at home. Finally, as one can expect, the male-female ratio is well balanced.

These variables are the common predetermined characteristics used in the literature. But most importantly, they are also variables that are observable by a school administrator (or, in general, by a social planner). One major difference with the literature is that we do not use previous achievement to classify students as high or low ability. Instead, we rely on psychological examination on intellectual giftedness (high ability), and special needs or age at test (low ability).

# 3 Setting

## 3.1 Empirical setting

Peer-effects studies or social interaction models are traditionally interested in the following coefficient $\gamma$:

$$y_{ic} = \alpha + \beta T_{ic} + \gamma \overline{T}_{(-i)c} + \delta \mathbf{X}_{ic} + \mu_c + \varepsilon_{ic}, \tag{1}$$

where $y_{ic}$ is the outcome of student $i$ in cell $c$, $T_{ic}$ is usually an individual type binary indicator for the "own" effect of interest, $\overline{T}_{(-i)c}$ is the proportion of peers of a given type within the cell of reference. The proportion of peers in the same cell is usually computed as the "Leave-Own-Out" (LoO) proportion $\frac{1}{N_c}\sum_{j\neq i \in c} T_{ic}$, with $N_c$ being the cell size. $\mathbf{X}_{ic}$ are other covariates that do not define types (for instance, the cell size). $\mu_c$ gives the fixed effects at the level of randomization, and $\varepsilon_{ic}$ is the idiosyncratic error term. The estimand of interest $\gamma$ is thus the difference of expected values of $y_{ic}$ under different proportion of peers $\overline{T}_{(-i)c} = \bar{t}'_{(-i)c}$ and $\overline{T}_{(-i)c} = \bar{t}''_{(-i)c}$ in cell $c$ given own types and other control variables. In other words, it is the average effect of a change in the proportion of peers with given characteristics within a cell. In this setting, researchers usually isolate one variable of interest $T$ and its corresponding LoO proportion $\overline{T}$. For instance, much research has been done on the effect of being assigned to high or low proportions of female classroom mates (where $\overline{T}_{(-i)c}$ is the proportion of female classmates) for male and female students on school performance or long-term educational prospects (see for instance Brenøe and Zölitz, 2020). The proportion of peers is usually a continuous variable, but it is sometimes modeled as a binary indicator for exposure to a certain type of peers. Empirical papers using identification within schools over time often also include time dummies and sometimes school-specific time trends. In practice, researchers estimate Equation (1) with a linear regression (linear-in-means model) and add polynomials of the LoO proportion to explore potential nonlinearities of the effect of $\overline{T}$. In most applications, $\mathbf{X}_{ic}$ also contains covariates about individual characteristics under the assumption that these characteristics might confound the investigated effect of peers.

We extend this single variable approach in two ways. It is in general not realistic to look at the effect of a shift in the proportion of $\overline{T}_{(-i)c}$ without looking at all the other peer effects that interact simultaneously to a shift in $\overline{T}_{(-i)c}$. Consequently, we consider models where (i) own types are interacted with each other (e.g., a student can simultaneously be a female and a nonnative student); (ii) where LoO proportions reflect these interactions (e.g., the proportion of female nonnative students in a cell); (iii) and where LoO proportions and own types are interacted (e.g., the effect of female peers might affect

nonnative students and native students differently). As an illustration, for a case with $n$ binary types, we end up with $2^n$ possible groups, thus $2^n$ LoO variables, and $2^n \times 2^n$ interactions between groups and their corresponding LoO proportions. If we add polynomials of the LoO proportions, we multiply again the number of variables by $2^n$. This leads to the second way we extend the traditional single-variable approach. As the number of potential predictors of $y_{i,c}$ becomes larger very fast (such that the number of predictors might become as large as the number of observations), we assume that only a subset of variables $S < p$ has an influence on the outcome variable. This assumption that the "true" model has only a (relatively) small number of nonzero parameters is known as the *sparsity assumption* in the ML literature.

Before we explain in more detail how we deal with these two extensions from an estimation perspective (in Section 3.3 and Section 3.4 below), we now turn our attention to the identification strategy.

## 3.2 Identification

The coefficient $\gamma$ is identified with the provision that three identifying assumptions hold: first, there must be no reflection problem caused by the fact that individual outcomes, peer outcomes, and students' and peer characteristics are determined simultaneously in the cell of interest (see Manski, 1993). To account for the reflection problem, we only use covariates defined before the assignment to schools or classrooms. That is, we use covariates defined prior to entering middle school.

Second, there must be no common and correlated unobserved shocks at the group level. To resolve this issue, we control for unobserved heterogeneity at either the school-by-track or the school-by-track-by-year levels. More precisely, we apply two different identification strategies. First, we exploit the natural variation in the cohort composition within school-tracks. This strategy rests on the assumption that small differences in cohort composition over time within the same school are unrelated to other factors determining academic performance. This approach has become the "gold standard" in the peer effects literature (e.g., Angrist and Lang, 2004; Bifulco, Fletcher, and Ross, 2011; Black, Devereaux, and Salvanes, 2013; Brenøe and Zölitz, 2020; Carrell and Hoekstra, 2010; Carrell, Hoekstra, and Kuka, 2018; Lavy, Paserman, and Schlosser, 2012; Lavy and Schlosser, 2011). However, spillover effects are more likely to emerge in the classroom, because the classroom is the policy-relevant peer group in education production (Lazear, 2001). Therefore, and as a second strategy, we also exploit variation between classrooms within the same school-track-years (e.g., Burke and Sass, 2013).

Third, there must be no endogenous peer selection into cells. Selection into cohorts would threaten the validity of our first identification strategy. For example, an idiosyncratic migration shock could affect the distribution of nonnative students in some cohorts. Selection into classrooms (within the same school-track-year) would threaten the validity of our second identification strategy. For example, this would be the case if a school principal strategically assigns nonnative students to classrooms or teachers. Even though we are estimating effects within school-track-year, this second approach is generally less plausible than the first strategy because of potential selection into classes. We conduct this second analysis anyway, and present balancing checks for both identification strategies.

To assess whether there is selection of students into cells, we conduct two different balancing tests. Conducting these tests is not trivial in our setting, because we must test for the selection of students with respect to five characteristics – and not only one, as in standard peer-effect studies. In addition, we are interested not only in the balance of the first two moments (mean and standard deviation of students' characteristics across cohorts and classrooms), but in the distribution of LoO proportions, as we investigate nonlinearities in the effects. In a first test, we conduct a randomization analysis as in Bifulco, Fletcher, and Ross (2011). We randomly draw observations at each level of randomization (school-track for cohorts, school-track-year for classrooms) 500 times with replacement and compare the randomly sampled distribution of student types within each level of treatment with the distribution we observe in the data. As can be seen in Table 2, the random distribution and the actual distribution are very similar in terms of first and second moments. When testing for differences in means with a $t$-test, we find no difference between the mean of the actual distribution and the mean of the random distribution. This is corroborated by the shape of the respective distributions presented in Figure A.1 and Figure A.2: randomized distributions of types fit the actual distributions well. All in all, these findings suggest that both identification strategies are plausible.

[Insert Table 2 here]

To detect potential selection into classrooms or into cohorts, we conduct further balancing checks by testing whether each peer effect variable (the mean of LoO variable) predicts individual baseline characteristics (gender, native speaker, relative age, special needs, and giftedness) conditional on school-track or school-track-year fixed effects (as in Guryan, Kroft, and Notowidigdo, 2009). We regress each mean of LoO variables on the other baseline characteristics, and we control for the relevant fixed effects (school-track-year for classroom identification and school-track for cohort identification) as well as for

the mean peer type at the level of randomization (this last control is the "correction" proposed by Guryan, Kroft, and Notowidigdo 2009). For five main characteristics, we regress $5^2 - 5 = 20$ regressions and report the distribution of p-values for the coefficient of the characteristics of interest.

[Insert Figure 1a and Figure 1b here]

The distribution of p-values for the balancing tests across classrooms is presented in Figure 1a. Two p-values out of 20 are under the significance threshold of 5% for the cohort identification strategy. For the classroom identification strategy, four p-values out of 20 are under the threshold of 5%, which indicates that some peer effects are predictive of individual characteristics in the classroom. These four significant coefficients are the coefficients of the share of older peers on the special need status (and vice-versa), and the coefficients of the share of nonnative peers on special need status (and vice-versa).

While the tests indicate no threat to internal validity for identification across cohorts over time, there are some reservations with identification between classrooms across schools. For the latter strategy, even though the distribution of types in the data are consistent with a randomly generated one, we suspect that a correlation exists between peers with SEN and older peers. This is partially expected, because many student with SEN either enter school one year later or repeat a grade in elementary school. To mitigate these concerns, all estimates "control" for both individual and cell characteristics.

### 3.3 Stable Selection

In this section, we explore the problem of finding the number of nonzero parameters that are in the "true" model. We want to know which variables are the most important predictors of school performance, and in which combination. We imagine a school official who wants to understand what variables are crucial determinants of academic performance in an inclusive educational setting. Since she observes a potentially high-dimensional set of student characteristics, she is likely interested only in the most relevant variables. In practice, she observes a set of $p$ student characteristics ("types") $t_j, j = 1, ..., p$ and their corresponding leave-own-out variables (LoO variables, or "peer effects") $\bar{t}_j, j = 1, ..., p$. From this set of types and peer effects variables, she would like to focus on the subset of variables $S \subseteq t_1, ..., t_p; \bar{t}_1, ..., \bar{t}_p$ that are the most informative of students' school performance. The larger $p$ is, the more relevant the knowledge about $S$ becomes for the school official, as a model with too many parameters becomes prac-

tically unusable.[8]

We implement the stable selection algorithm proposed by Meinshausen and Bühlmann (2010) and refined by Shah and Samworth (2013) to restrict the set of relevant variables $S$ to a low-dimensional, and practically usable, set of influential characteristics. For notation, we follow Hofner, Boccuto, and Göker (2015). We draw $B$ random subsets of the sample of size $\lfloor n/2 \rfloor$ and, on each subsample $b = 1, ..., B$, we fit the lasso statistical learner that selects a set of features of maximal size $q$. In brief, lasso (or "least absolute shrinkage and selection operator", or "$l1$-penalized regression") is a widely used ML method based on linear regression that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of linear models with many variables (or "predictors"). The main tuning parameter for the lasso is $\lambda$, which gives the amount of regularization (if $\lambda = 0$, the model is identical to OLS without penalization, and the larger $\lambda$, the more coefficients are set to 0). The optimal value of the parameter $\lambda$ is usually defined *ex ante*, or with cross-validation procedures. In our case, for each random subsample, we run the lasso at a particular value of the shrinking parameter $\lambda \in \Lambda$, where $\Lambda$ is defined as the candidate set of $\lambda$ values between the highest value of $\lambda$ such that no coefficients are selected, and the smallest value of $\lambda$ such that $q$ coefficients are selected. The set of selected variables per subsample $b$ and per value of the regularization parameter $\lambda$ is $\hat{S}^b_\lambda$. We then compute the empirical probability for a variable to be selected:

$$\hat{\pi}^\lambda_j = \frac{1}{B} \sum_{b=1}^{B} I_{j \in \hat{S}^b_\lambda}, \tag{2}$$

where $I$ is the indicator function that takes value one if the variable $j$ was selected in the subsample $\hat{S}^b_\lambda$. This gives the stability path, which is the probability for each variable to be selected at a given value of $\lambda$ when randomly resampled from the data. Finally, we select all the predictors that are selected with a selection probability of at least $\pi^*$, which is a pre-specified threshold value. By doing so, we end up with a set of stable variables $\hat{S}_{stable} = \{ j : \max_{\lambda \in \Lambda} \hat{\pi}^\lambda_j \geq \pi^* \}$.[9] Meinshausen and Bühlmann (2010) show that results do not depend strongly on the value of the regularization parameter $\lambda$.

As the lasso is a shrinkage and selection method for linear regression, it allows us to directly relate

[8]For simplicity, we focus on individual types and their corresponding leave-own-out variables. However, other variables can influence students' performance, such as group size.

[9]This selection procedure ensures that the false positives rate $V$, i.e., the rate of wrong selections, has an upper bound given by $E(V) \leq \frac{1}{(2\pi^* - 1)} \frac{q^2}{p}$. The expected number of wrong selections is lower when the number of chosen features $q$ is lower and when the threshold $\pi^*$ is higher. For more details on the error rate bound, see Meinshausen and Bühlmann (2010) and Shah and Samworth (2013).

our results to the existing literature on peer effects that uses linear-in-means models. However, since the set of variables $S$ might contain many interactions between types, and between types and LoO variables, we need to use a version of lasso which is able to account for both strong hierarchy and weak hierarchy. Strong hierarchy assumes that an interaction can be part of the true underlying model only if both its main effects are also part of the true model, whereas weak hierarchy assumes that an interaction can be part of the true model as long as one of its main effects are part of the true model. The hierarchical group lasso developed by Lim and Hastie (2015) is able to catch interactions and main effects that obey strong hierarchy as well as weak hierarchy. In contrast, the traditional lasso might select only interactions without their main effects, which might not make sense in our settings. The hierarchical group lasso is based on the group lasso (Yuan and Lin, 2006), which conducts variable selection by setting groups of variables to zero. It essentially defines main effects and interactions as belonging to the same group of variables, and performs variable selection on these groups. We adapt the stable selection algorithm in order to accommodate the hierarchical group lasso.[10]

### 3.4 High-Dimensional Peer Effects

In this section, we estimate high-dimensional peer effects with flexible learners based on ML in order to discover heterogeneities and nonlinearities in effects. We extend the model presented in Equation (1): the test score for students with a vector of characteristics $\mathbf{T}_{ic}$ is estimated under particular corresponding $\overline{\mathbf{T}}_{(-i)c}$ peer proportions (LoO variables defined at the cohort or classroom level $c$), and other covariates of interest $\mathbf{X}_{ic}$. By definition, the vector of individual types and the vector of corresponding LoO proportions have the same length. We therefore want to flexibly estimate the following function:

$$y_{ic} = g(\mathbf{T}_{ic}, \overline{\mathbf{T}}_{(-i)c}, \mathbf{X}_{ic}, \zeta(\mathbf{T}_{ic}, \overline{\mathbf{T}}_{(-i)c}, \mathbf{X}_{ic})), \tag{3}$$

where $\zeta(\mathbf{T}_{ic}, \overline{\mathbf{T}}_{(-i)c}, \mathbf{X}_{ic})$ represents a (high-dimensional) vector of covariates created from the interactions of all the elements of $\mathbf{T}_{ic}, \overline{\mathbf{T}}_{(-i)c}, \mathbf{X}_{ic}$ with each other (and also possibly including interactions of higher levels as well as polynomials). Moreover, like in the stable selection exercise, we assume sparsity, i.e. that only a subset of predictors among the predictors in function $g()$ are nonzero in the "true" model. We estimate this function by means of flexible ML learners. In our setting, since the assignment to cohorts within school-tracks, or to classrooms within school-track-years, is random, we are able to assume that there is no selection into cells. The traditional peer effect research is interested in the following

[10] All code files can be found on `https://github.com/ASallin/hdpx`.

estimand, which conceptually corresponds to the coefficient $\gamma$ in Equation (1):

$$\Delta_{\overline{t'}_{c,k},\overline{t''}_{c,k}} \quad = \quad E[y_{ic}|\overline{T}_{(-i)c,k} = \overline{t'}_{c,k},...] - E[y_{ic}|\overline{T}_{(-i)c,k} = \overline{t''}_{c,k},...],$$

where $\overline{T}_{(-i)c,k}$ is now the scalar representing the $k$th element of $\overline{\mathbf{T}}_{(-i)c}$. In other words, $\Delta_{\overline{t'}_{c,k},\overline{t''}_{c,k}}$ is the marginal effect of a change in the $k$th LoO proportion from $\overline{t'}_{c,k}$ to $\overline{t''}_{c,k}$ at point $\overline{T}_{(-i)c,k} = \overline{t'}_{c,k}$. Note that the difference between $\overline{t'}_{c,k}$ and $\overline{t''}_{c,k}$ is usually picked to be a meaningful variation in the peer proportion of a certain type $k$ (such as the change in $\overline{T}_{(-i)c,k}$ that represents one additional student per classroom). As we do not expect the effect to be linear, this difference changes with the particular value of $\bar{t}_{c,k}$, of $\overline{t'}_{c,k}$, as well as with all the other variables $\bar{t}_{c,k}$ and $\overline{t'}_{c,k}$ are interacted with. The magnitude $\Delta_{\overline{t'}_{c,k},\overline{t''}_{c,k}}$ can be understood as the average treatment effect (ATE) of a shift in $\overline{T}_{(-i)c,k}$.

Moreover, we will be interested in the heterogeneous effects of a shift in LoO proportions for the $l$th element of $\mathbf{T}_{ic}$, i.e., for type $l$:

$$\Gamma_{\overline{t'}_{c,k},\overline{t''}_{c,k},t_{c,l}} \quad = \quad E[y_{ic}|\overline{T}_{(-i)c,k} = \overline{t'}_{c,k}, T_{ic,l} = t_{c,l},...] - E[y_{ic}|\overline{T}_{(-i)c,k} = \overline{t''}_{c,k}, T_{ic,l} = t_{c,l},...]$$

where $k = l$ or $k \neq l$. For instance, the effect of having more gender peers is investigated for female ($T_{ic,l} = 1$) and male ($T_{ic,l} = 0$) students separately. These conditional effects can be investigated at the level of category types (for instance, across gender), but can also be investigated at the quasi-individual level (for granular types, such as, for instance, the effects for nonnative male students with SEN, and for nonnative male students without SEN).

To estimate the function $g()$, we develop the following procedure. In step (1), we demean all variables in our dataset at the level of randomization (school-track or school-track-year). This allows us to make predictions and conclusions that account for unobserved cell effects that would confound our outcome (conceptually, this is similar to a "fixed effects" procedure). In step (2), we conduct clustered $k$-fold cross-fitting, where "clusters" in this context stand for the cells forming the level of randomization (particular school-track cells for the cohort identification or particular school-track-year cells for the classroom identification). In step (2), we (2a) first randomly assign each cluster to $k$ different groups (or "folds"), and (2b) within each fold, we randomly draw 80% of observations. In step (3), we use the drawn observations in $k-1$ folds to train a ML learner that predicts the outcome of interest.[11] In step (4),

[11] As ML learners, we use both hierarchical group lasso and random forest. To combine the predictions of these two different models, we program an ensemble learner that combines all predictions into one vector of fitted values in such a was that the RMSE is minimized.

we fit our trained model on the left-out $k$th fold. This is known as "out-of-bag" prediction. In step (5), we repeat step (2) to (4) $k$ times such that we have out-of-bag predictions for all $k$ folds. Finally, in step (6), we repeat step (2) to (5) $M$ times and obtain a matrix of $n \times M$ fitted values, where $n$ is the number of observations in our sample. These $M$ repetitions are important for inference: similar to bootstrapping in the case of linear regression, we obtain a distribution of fitted values per observation, and estimate standard deviations for fitted values.[12]

This clustered cross-fitting procedure aims at minimizing the risk of overfitting while producing externally valid results. ML methods could easily fit the data perfectly, and thus our predictions would not be informative outside of our data. In fact, we want our results to generalize outside of the school-track cells or outside of the school-track-year that form our sample, such that our conclusions would also apply to students that are in cohorts or classrooms we do not observe in our dataset. Since we train our ML learners on 80% of observations in the $k-1$ folds, and that we predict our model on "out-of-bag" observations in the $k$th fold, we never use the same observation and the same clusters to train the model and to obtain predictions. This ensures that our models are robust to differences in clusters (see, as an inspiration, Athey and Wager, 2019). This procedure is necessary to capture effects for students whose types are relatively rare (e.g., gifted students).

Finally, in order to summarize and represent $\Delta_{\overline{t'}_{c,k},\overline{t''}_{c,k}}$ and $\Gamma_{\overline{t'}_{c,k},\overline{t''}_{c,k},t_{c,l}}$ at different LoO proportions points for the whole population or for subpopulations of interest, we use simple group means at different points of $\Delta_{\overline{t'}_{c,k},\overline{t''}_{c,k}}$ and $\Gamma_{\overline{t'}_{c,k},\overline{t''}_{c,k},t_{c,l}}$. To show how the effects vary at each value of the LoO proportion, we estimate kernel regressions in which we flexibly regress the predicted outcome on a peer composition of interest $\overline{T}_{ic} = \bar{t}_{ic}$. These kernel representations give the marginal peer effect of a given peer type on peers' academic performance at each $\bar{t}_{c,k}$ point.

# 4 Results

## 4.1 Stable Selection

**True model scenarios.** We conduct stable selection on academic performance with $B$ =200 draws and a threshold value of $\pi^*$ =0.75, which means that we keep variables that have a probability of 0.75 to be

[12]Note that, in comparison to classical bootstrapping, 50 repetitions might seem like a low number of repetitions. Our constraints in the number of repetition purely stem from our computational limitations, as training ML algorithms with the clustered $k$-fold cross-fitting procedure is computationally intense.

selected by the algorithm. Since we are *a priori* agnostic about the presence of interaction terms (and their complexity level) in the "true" model, we present three scenarios. First, we assume that the true underlying model contains only main effects, i.e., the effect of own main types and the effect of peers of each of these own types. This setting corresponds to the setting adopted in most traditional peer-effect studies. Second, we assume that the true model flexibly comprises of interactions of degree two between types, between LoO variables, and between types and LoO variables. Finally, we look at types at the finest degree of interaction, i.e., variables that exhaust the field of possible interactions between types. For instance, our five main types, which depict ten main categories (female and male, older and non-older, gifted and non-gifted, etc.), result in $2^5 = 32$ types. These 32 types give the finest degree of type representation, and each observation belongs to one and only one type. We interact each of these 32 types with their corresponding 32 LoO variables.[13]

**Results of the stable selection.** We present results of the stable selection exercise at the cohort level in Figure 2 without interactions, Figure B.1 with interactions, and Figure 4 of the Appendix for fully interacted types. For the classroom level analysis, we present results in Figure 3 without interactions, Figure B.2 with interactions, and Figure 5 of the Appendix for fully interacted types. In each graph, we present, on the left-hand side panel, the selection probability for the selected variables. This probability represents the probability for the variable to be selected by the hierarchical group lasso in 200 draws averaged across all values of the parameter $\lambda$.[14] We want to emphasize the fact that this analysis is *qualitative* in nature: we want to learn which variables are important from a predictive point of view. However, in order to have an idea of the effect size and to relate our results to the literature, we report, on the right-hand side panel, the corresponding OLS regression coefficients of all selected features (with fixed-effects at the school-track level for the cohort analysis, and at the school-track-year level for the classroom analysis). All selected predictors are additively included in a single regression, and, for this reason, effects must be interpreted in a *ceteris paribus* way, i.e., the effect of a given variable when all

[13] From these 32 types, we exclude types that are always zero, as well as types that cover less than 1% of the population. In case of variables that are collinear or highly correlated, the lasso will likely pick one variable over the other by chance. We do not think it is a problem in this setting for the following reasons: first, the main characteristics are not collinear. Second, a school administrator or a policy maker would not find real value in learning about all highly correlated variables. For this reason, we remove *ex ante* collinear predictors when we conduct the analysis on full types. Finally, because we decompose each combination of types to their most granular level, the fully interacted model does not contain main effects (such as, for instance, the effect of "only" being a male student). Including all main effects, interacted effects of depth 2, interacted effects of depth 3, and so on, would end up in a model with too many (highly correlated) predictors.

[14] The path for values of $\lambda$ is defined in such a way that the number of variables selected is under the chosen number of variables in the true model $q$. Although varying the values of $q$ does not influence the results, we arbitrarily choose $q = 6$ for models without interactions, $q = 40$ for models with interactions, and $q = 25$ for the fully interacted models. We pick a lower value for $q$ for the fully interacted model as we already have "thrown out" highly correlated predictors.

the others are held constant.[15]

The analysis without interactions in Figures 2 and 3 reveals insightful results. At the cohort level, only the five own types are selected and no peer effects are present. This means that peers' influence at the cohort level does not predict academic performance. At the classroom level, four out of the five selected variables are own types. These results confirm the literature that own characteristics are more important than peer effects (see the discussion by Angrist, 2014).

[Insert Figure 2 and Figure 3 here]

Interestingly, the impact of peer students with SEN is highly predictive of academic performance in the classroom environment, which confirms the analysis of Balestra, Eugster, and Liebert (forthcoming) in the same setting. This is even more noteworthy since the own SEN status is the only own type not selected in the classroom setting. This is likely due to the fact that the own SEN status does not necessarily predict own achievement: on the one hand, there is heterogeneity among all students with SEN, and many of them are diagnosed with disabilities or issues that are not related to school performance. On the other hand, the main effect of SEN may go through another variable (e.g., nonnative or gender). All in all, the presence of students with SEN in the classroom impacts the school performance of their classmates as they may disrupt learning or need additional teachers' attention. In terms of effect size, all types are negatively correlated with academic performance except for the giftedness status.

Results of variable selection when allowing for interactions of level two offer an even more comprehensive picture of which variables are the most relevant in a peer-effect setting. As presented in more detail in Appendix B (Figures B.1 and B.2), the effect of older peers and the effect of peers with SEN are the dominating spillovers at the cohort level. Moreover, these two peer effects are heterogeneous: both effects are interacted, meaning that the effect of older peers changes as a function of the proportion of peers with SEN in the cohort.

At the classroom level, the five main types are selected. The first dominating peer effects are spillovers from peers with SEN. The algorithm selects, in 100% of cases, peer effects from students with SEN on other students with SEN, and peer effects from students with SEN on nonnatives. However, and surprisingly, the main effect of peers with SEN is not selected: this interesting case of weak

[15] A cautionary note on the interpretation of p-values in the graph: it is important to remember that the p-values from the post-selection regressions cannot be interpreted in the traditional sense, since the variables have been selected in a first step. Some of the coefficients may not be statistically significant even though their variable has been selected.

hierarchy means that the main effect of classmates with SEN is not part of the "true" model. The second dominating effects in the classroom are effects from older peers (see also Bietenbeck, 2020): older peers have a negative impact on their peers, and this impact is interacted with the classroom size and with the effect of female peers. Finally, some variables are interacted with the classroom size: the effect of peers with SEN, the effect of older peers, and the own SEN status. The influence of classroom size for students who are more likely to fall behind is anything but surprising: their academic success is more likely to depend on the availability of teaching resources and individual teacher attention.

[Insert Figure 4 and Figure 5 here]

Turning our attention to fully interacted models, we analyze heterogeneities at the most granular level possible. The analysis of fully interacted models at the cohort level as presented in Figure 4 shows that 13 variables are selected, among which seven variables are peer effects. Both genders are selected, either alone or in combination with other characteristics.[16] Among these seven variables, two of them reveal peer effects from male peers with SEN on female students or low-achieving female students, and two of them reveal peer effects from male peers on low-achieving female students. The other three are peer effects from female students on other female students. These findings reinforce the conclusion that peer effects are mostly due to male peers on female students who are more at risk of under-performing academically (female students with SEN or older, or nonnative).

At the classroom level, Figure 5 shows that 13 variables are selected, among which five variables are peer effects. Again, most important peer effects are generated by students with SEN who are male and nonnative, and by students with SEN who are male, nonnative, and older. These peer effects can be related to the "disruption" hypothesis put forward by many empirical papers (e.g., Balestra, Eugster, and Liebert, forthcoming; Bietenbeck, 2020; Bifulco, Fletcher, and Ross, 2011; Carrell, Hoekstra, and Kuka, 2018; Figlio, 2007; Hanushek, Kain, and Rivkin, 2009). In addition, female peers with no other observed characteristics generate positive spillovers, whereas male peers with no other observed characteristics generate positive spillovers mostly on male classmates. These results, and the fact that both effects are positive, are in line with Bertrand and Pan (2013) and Lavy and Schlosser (2011). The characteristic of being a gifted male is selected, whereas being a gifted female student is not. This reflects the findings by Balestra, Sallin, and Wolter (forthcoming) in a very similar setting. Finally, spillover from female peers

[16]Note that the variables "male" or "female" correspond to male and female students who are not older, not nonnative, not gifted, and without SEN.

both on male students and on other female students are selected, in contrast to the results for the cohort level. As our results and our simulation exercise below suggest, female peers are a major (positive) force in the classroom, independently of the other characteristics also associated with a female student.

In conclusion, the stable selection exercise offers valuable qualitative information on how spillovers interact in school groups. It shows that spillover effects are dominated by the effect from students with SEN and from low-achieving students, more particularly from low-achieving male students with SEN. Special needs and relative achievement are therefore important factors to take in consideration, especially in classrooms. Moreover, and in general, we observe that peer effects are more "diluted" at the cohort level, which is expected as peers exert most of their influence in smaller groups. This is consistent with Burke and Sass (2013) who discuss how peer effects change between cohort and class level. From a methodological point of view, the stable selection exercise confirms our knowledge that individual effects are more important than peer effects. Therefore, it is important to "control" for own types when estimating spillover effects. Finally, it shows that effects and spillover effects are most likely heterogeneous across types, and depend heavily on the group composition.

### 4.2 High-Dimensional Models

We estimate high-dimensional models using ML algorithm as presented in Section 3.4, both at the cohort and at the classroom levels. All variables are demeaned at the level of randomization, so coefficients can be interpreted as differences from the mean of the randomization cell. Since the estimated spillover functions account for many parameters, they can be used to represent various effects of interest. In what follows, we represent the functions in a "classical", non-interacted way, keeping in mind that the fitted values were estimated using all interactions and parameters selected in the previous section. To achieve a graspable sense of the effects' magnitude and sign, the effects are summarized by a kernel regression that takes the predicted values from the ML algorithm as outcome and the LoO variable of interest as the predictor.[17]

**Results at the cohort level.** We first discuss the high-dimensional functions for the cohort identification strategy, which are presented in Figure 6. Each subgraph presents the effect of peers of a given type (for instance, share of female peers) both on peers of the same type in red (female students), and on peers

[17] Kernel bandwidths are found by cross-validation. Given our algorithms, we conduct inference across the the $M$ predictions. This means that, for kernel regressions, we estimate $M$ kernel regressions, and we conservatively report, for each predicted LoO point, the upper confidence interval across the $M$ predictions, and the lower confidence interval across the $M$ predictions. This explains the peculiar shape of confidence intervals in reported figures.

of the other type in blue (male students).

Figure 6a shows the peer effects from older peers. Students in cohorts with more older peers are negatively affected, and this is true both for students who are older and for students who are not older (the two regression lines have the same slopes). A similar pattern can be observed for peers with SEN in Figure 6b. The representation of spillover effects for gifted students in Figure 6c shows that gifted students perform on average 0.5 standard deviations higher than nongifted students, and that nongifted peers benefit from the presence of gifted peers in their cohort. Note that the curves for gifted students do not cover as much support as other variables, as the prevalence of gifted peers is on average very low. Nonnative peers at the cohort level seem not to affect their peers. Finally, having more female peers does not seem to affect either male students nor female students. This last result is important, as it shows that peer effects originating from gender do not happen at the cohort level.

[Insert Figure 6 and Figure 7 here]

**Results at the classroom level.** At the classroom level, peer effects appear to have a larger impact on school performance. This is expected, as the classroom environment is the environment in which peers interact the most. Figure 7a shows that a higher proportion of low-achieving, older peers in the classroom has a negative influence on peers, both low-achieving and not. The picture is, again, similar in the case of peers with SEN. In both cases, the estimated spillover functions are downward sloping, which suggests that classroom environments with lower segregation provide higher chances of academic success for both students with SEN and students without SEN. The share of nonnatives in the classroom has no negative impact for lower proportions of nonnative peers, but peer effects from nonnative peers are (largely) negative for classrooms with a higher degree of nonnative students. It is important to note that effects generated from older students, students with SEN, and nonnative students, are downward sloping and not constant in the share of peers within classrooms. In fact, the negative marginal effects of adding additional older peers, peers with SEN, or nonnative peers in classrooms increase in the proportion of these peers. This means that the more older peers, peers with SEN, or nonnative peers in the classroom, the more negative their impact on other students and on themselves become. For instance, the marginal effect of having one percentage point higher proportion of students with SEN is around -0.5 percentage points of a test score standard deviation on the students without SEN when computed at the average proportion of students with SEN in the school-track-year. However, this effect is -1.5 percentage points when computed in classrooms with a proportion of students with SEN which is 0.2 percentage points

higher than the average. Finally, the share of female peers has a positive impact on academic performance for both female and male students. This is different from results found at the cohort level, which suggests that female peers exert positive spillovers in classrooms rather than cohorts.

In an additional step, we investigate heterogeneities of spillover effects across gender. Many studies investigate heterogeneous peer effects across gender to understand, for example, the gender gap in school performance (Fryer and Levitt, 2010), career choices (e.g., Brenøe and Zölitz, 2020), and preferences (Niederle and Vesterlund, 2010). Our framework allows us to investigate such heterogeneities in a systematic way. To have a better idea of how peers affect female and male students, we summarize our estimations of gender heterogeneities with kernel smoothers in Figures 8 and 9.

[Insert Figure 8 and Figure 9 here]

We find interesting overall results: first, at the cohort level, there does not seem to be substantial effect heterogeneities along gender alone. For all four main peer categories, nonlinear effects for female and male students have similar slopes. Second, at the classroom level, we observe that male students tend to be more negatively impacted by a higher share of peers with negative influences (older peers, peers with SEN, and nonnative peers). Above a residual classroom proportion of 0.2 in peers with negative influences, the gender gap tends to decrease, as the negative slopes for female students are smaller than the negative slopes for male students. Thus, the effect of disruptive peers seems to affect male students the most, which is consistent with the existing literature (Bertrand and Pan, 2013). These results are also valuable as they serve as cautionary tales for research about the gender gap in classroom: linear models might detect gender gaps because of extreme values in peer proportions.

**Ideal school environments.** Finally, we use the estimated functions to investigate ideal school environments in a partial-equilibrium setting. We look at the characteristics and peer composition of cohorts and classrooms that have the highest average predicted test score. Table 3 shows the peer composition of these cohorts and classrooms with bootstrapped standard errors. Our results show that, almost trivially, cohorts or classrooms with the highest average predicted test scores are those with the lower proportion of peers with negative effects and higher proportion of peers with positive effects. Classrooms with the lowest aggregated test scores have a proportion of older peers which is 10 percentage points lower than the average proportion in their school-track-year cells, 14 percentage points lower proportion of students with SEN, 17.5 percentage points lower proportion of nonnative students, and 5.5 percentage points less

female students. However, these classrooms have around 2 percentage points higher proportion of gifted peers. One interesting finding is the role of group size, as the best environments are always smaller than the average size in the randomization pool. Best classroom environments have, on average, 0.7 less students than the average classroom at the school-track-year cell. This results corroborates the extensive literature on class size (Lazear, 2001).

[Insert Table 3 here]

# 5 Policy Counterfactuals

Probing further, we turn to an analysis of interesting policy counterfactuals and conduct the following exercise in the spirit of, for example, Graham et al. (2020). Using our high-dimensional peer effect function estimates, we want to understand how spillovers affect students on an aggregate level, i.e., when the full population of students is considered. To perform this analysis, we look at general equilibrium effects of marginally varying the classroom compositions while keeping the student population constant. Practically, the school administrator decides to remove, from each classroom, one student of a given type, and to reallocate these students into a single classroom. Doing so, one classroom per school becomes a "cluster" of students of given type. This counterfactual scheme has the advantage of being feasible and easily implementable for the school administrator. In addition, this approach would limit concerns of teachers' adaptation in teaching technology and, most importantly, it would not rely on extrapolations for estimation – as all classroom compositions are actually observed in the data. These simulated classroom compositions represent counterfactual allocations whose aggregated outcomes can be evaluated with our flexible ML spillover functions. We therefore compare the overall average academic performance under a random classroom allocation and under the simulated allocation. For completeness, we conduct a further simulation exercise when we fully segregate along types (instead of only removing one student per classroom) in Appendix C.

We impose the following three constraints that plausibly mimic the situation of a school administrator who has only a given number of teachers and classrooms at her disposal, but who is interested in implementing different classrooms compositions. First, the pool of students allocated to classrooms is randomly drawn from the population and has a fixed size. This illustrates natural variations in the pool of newly enrolled students a school administrator faces every school year. The school administrator must

allocate all these students to a classroom. Second, classrooms have equal size. This constraint is set to facilitate computations, and could be easily relaxed. Third, the objective function is the maximization of average test scores at the school level, implicitly assuming an equal welfare weight for all students (each student is considered equally in the welfare function independently of her characteristics).

To estimate effects of different classroom allocation schemes, we are interested in the following parameter of interest: the *average counterfactual effect ACE*:[18]

$$\widehat{ACE} = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{y}_{ic}^{\text{segr.}} - \hat{y}_{ic}^{\text{random}}\right], \tag{4}$$

where $\hat{y}_{ic}$ is the predicted outcome from the estimated function $g()$ under a more segregated allocation or a random classroom allocation. We are also interested in the *conditional average counterfactual effect CACE*, which is the average counterfactual effect for students with varying baseline types $t_c$. For instance, we are interested in whether students with SEN would be affected differently from increased segregation than nonnative students. The *CACE* is:

$$\widehat{CACE} = \frac{1}{\sum_{i=1}^{N}\mathbf{1}(T_{ic}=t_c)}\sum_{i=1}^{N}\mathbf{1}(T_{ic}=t_c)\left[\hat{y}_{ic}^{\text{segr.}} - \hat{y}_{ic}^{\text{random}}\right]. \tag{5}$$

Finally, we also compute the (average) classroom Gini coefficient to measure equality across the counterfactual allocation regimes.

To estimate both the *ACE* and the *CACE*, we proceed as follows. We randomly draw a pool of 100 students out of the main sample (our "school"), and we randomly create 5 classrooms of 20 students each. This is the random classroom allocation setting. The number of 20 students is chosen to ease computations, and corresponds roughly to the average classroom size in our setting. In a second step, we remove one student of a particular type (e.g., one student with SEN) from each of the four classrooms and "cluster" these four students in the fifth classroom. We then randomly remove four students of other types (e.g., students without SEN) from the fifth classroom and randomly allocate each of them to the four remaining classrooms. This keeps the classroom size constant. This setting is our "manipulated" allocation setting. For each student and for both settings, we generate the LoO variable at the classroom level along the five main characteristics. To obtain the individual predicted values, we match our randomly drawn observations with their nearest neighbors in the original sample, and we use our predictions

[18]The *ACE* directly reflects the *average reallocation effect ARE* of Graham et al. (2020). However, to accentuate the fact that we are not doing a *reallocation* exercise, we rename it as the *counterfactual* effect.

from our estimated $g()$ function. We conduct this exercise for 500 random drawn, and for each of the $M$ predicted values. We compute bootstrapped confidence intervals (as we have, for each random sample draw, $M$ predicted values). Given the number of random draws, the composition of classrooms under the random allocation corresponds, on average, to a balanced allocation of types to classrooms. We finally compare the *ACE* and *CACE* as the difference between the "random" classroom allocation scheme, and the "manipulated" classroom allocation scheme.

**Results for the *ACE*.** Results for the *ACE* are presented in Table 4. The picture is clear: the removal and clustering of students with given types has, on average negative but very small impact on the overall aggregated academic performance. For instance, settings in which one older peer is removed per classrooms have an aggregated test score performance which is 0.3 percentage points of a test score standard deviation lower than the setting in which students are randomly allocated to classrooms. The explanation for these very small differences is simply that the gains for students in classrooms without clusters is offset by the losses endured by students in classrooms with clusters. Interestingly, clustering nonnative speakers leads to a slightly positive improvement in aggregated score. All segregated settings slightly reduce the Gini coefficient, which means that, as expected, clustering leads to classrooms that are less homogeneous in types, and thus increases inequality. From a policy perspective, random allocation of students to classrooms seems to have higher aggregated outcomes than clustering students according to their types.

[Insert Table 4 here]

**Results for the *CACE*.** It is also possible to know more about the "losers" and "winners" of clustering students of given characteristics in one classroom. We look at the variations in predicted aggregate test scores for each type of student (as defined by the major five types, even though we could look at more granular types). Figures 10, 11, and 12 present the gains and losses for students of all categories when one older student (Figure 10a.), when one student with SEN (Figure 10b.), when one nonnative student (Figure 11a.), when one female student (Figure 11b.), and when one gifted student (Figure 12) per classroom are removed from their classrooms and clustered in one classroom. Bars in the graphs represent differences in group test score averages under the different counterfactual regimes for each particular type of students .

[Insert Figure 10 and Figure 11 here]

What happens when we cluster students of a given type into one classroom? For all spillovers that have a negative effect on classmates (spillovers from nonnative students, older students with SEN, and nonnative students), results follow the same pattern. Everyone in the classroom that contains the cluster is found to be harmed by the marginal increase in segregation (blue bars in the graph), and this negative impact is always larger than the gains for those who are kept in classrooms without clusters (red bars in the graphs). For instance, in Figure 10a, we see that all students, but especially male students, nonnative students, and students with SEN suffer when allocated to a classroom with a cluster of older students. However, all the other students in classrooms where one older student was removed are benefiting from the reallocation. In the case of the segregation of older peers, the average losses of nonnative students in clustered classrooms are as high as five times the average gains for those who happen to stay in the other classrooms. As we saw with the *ACE*, clustering nonnative students is, according to our results, the only classroom manipulation that has no negative impact (the blue bars are as large as the red ones for the subpopulations of native and nonnative students). The *CACE* helps us understand why: nonnative students on average benefit from being clustered in classrooms, while the natives kept in the classrooms without clusters are only slightly harmed.

Interestingly, the creation of gender clusters generates positive outcomes for students in the classroom with a cluster of female students. However, the gains for these students are balanced out by the losses for the other students who have more male peers in their classrooms. From a society perspective, gender-mixed education is the best solution in terms of maximizing aggregated test scores. These findings corroborate natural experiments exploiting segregation along gender in schools and in tertiary education (e.g., Eisenkopf et al., 2015; Pregaldini, Backes-Gellner, and Eisenkopf, 2020). The only category of students that does not benefit as much from gender-homogeneous environments are nonnative students. This is even more visible when we simulate full gender segregation in the appendix. We can only provide speculative interpretation of this: nonnative students in Switzerland mostly come from male-dominated cultures. Gender segregation might exacerbate male-dominated competitive behaviors.

Finally, clustering gifted students (Figure 12) generates positive spillovers for students in the classroom with a cluster of gifted students. Students who benefit the most from being in the classroom with a cluster of gifted students are male students, normal-age students, students without SEN, and native students. They benefit of a "boost" in predicted test scores of around 2.75 percentage points on average.

These results are in line with the findings of Balestra, Sallin, and Wolter (forthcoming), especially in the fact that low-achieving students (older students) do not react at all to a reallocation of gifted students. This framework also allows us to measure the effect of clustering gifted students on gifted students themselves: as the population of identified gifted students is very small, the effects of reallocating gifted students are zero and imprecisely estimated for gifted students.

[Insert Figure 12 here]

What are the main conclusions of this counterfactual exercise? First of all, all our results strongly suggest that attempts at making classroom more homogeneous in terms of types is not a good idea to improve aggregated test scores. Already "minimal" policies such as removing one student of a given type per classroom have negative impacts on the academic performance of the whole. This holds when we incorporate nonlinearities and full heterogeneity in effects. Even though clustering has a positive impact on students in classrooms where peers with negative influence are removed, this positive impact does not, on average, exceed the harm done on the students assigned to the classroom with cluster. Second, we show that inclusion (i.e. having classrooms that are mixed in types) decreases overall inequality. If we think of education as a public good, and the main mission of public schooling is to give anyone equal chances, having classrooms that are as heterogeneous in types as the population seems to be a good step in this direction.

These simulations are valuable as they give us a sense of the forces at play in classrooms from an aggregated perspective. Although inclusion and segregation are widely debated in schools, we show that classrooms that are balanced in terms of students' characteristics are Pareto-efficient, and we are not able to identify other allocation schemes that would be Pareto-improving. Of course, our conclusions hinge on four limiting assumptions. First, we have so far ignored group size effects. Second, we have not considered interactions of types (as in our stable selection exercise), which could refine our understanding of which population of students suffer the most from clustering. Third, we give each student a similar welfare weight in the objective function. If, for instance, schooling would be shown to have higher returns for gifted students, we might want to give them a higher weight. This is an ethical debate we are not willing to address in this study. Finally, we are only measuring test scores, and are ignorant about other outcomes that are influenced by classroom composition (such as psychological well-being, stress, etc.)

# 6 Conclusion

This study aims at giving a more realistic understanding of spillovers among students, and integrates two limitations that were not sufficiently taken into account by the existing literature. First, spillovers have heterogeneous impacts on individuals with different characteristics. Second, there are *at least* as many spillovers as there are types of students. To account for these two elements in our analysis, we build on the assumption that "true" spillover effects are nonlinear and high-dimensional, and we develop a general empirical approach that systematically considers nonlinearities and high-dimensionality of spillover functions using ML algorithms. More precisely, this study aims at discovering which spillovers influence academic performance the most, at learning about the way spillovers and individual characteristics interact, and finally at providing policy makers and school administrators with insights into (optimal) classroom compositions.

We run stable selection procedures to discover what are the spillover effects and other effects that influence students' test scores the most. We find that students' own characteristics are the most important predictors of their academic success, rather than students' peers. We also reveal that substantial heterogeneity hides between main effects, and that selected spillover effects are dominated by the effect from peers with special needs and from low-achieving peers. In a subsequent step, we design a flexible estimation procedure, and we find interesting heterogeneities in effects. Most importantly, we find that effects generated from older students, students with SEN, and nonnative students, are downward sloping and not constant in the share of peers within classrooms. Finally, we conduct simulations to investigate what would happen to a given student if he or she were put in a different classroom environment. We find that marginally increasing segregation has different impacts depending on which students are segregated. Our results clearly show that creating clusters of students with SEN has large negative impact on the aggregated school performance.

This study is a first step towards a more nuanced comprehension of peer effects in inclusive educational settings using new estimation techniques. Our main policy message is that even small manipulations in the classroom allocation of students have important consequences on students' academic performance. Therefore, a classroom allocation that "spreads" students of all types across classrooms is likely to mitigate the negative effects of clustering and, at the same time, "distributing" the peers that generate positive spillovers as much as possible (like, for instance, female students). Finally, while broadening our understanding of how to best serve students in inclusive school settings through mean-

ingful and policy-relevant simulation exercises, this study invites further research and more complete policy analysis regarding the inclusion of teachers and school resources.

## References


Angrist, Joshua. 2014. "The perils of peer effects." *Labour Economics* 30:98–108.

Angrist, Joshua and Kevin Lang. 2004. "Does school integration generate peer effects? Evidence from Boston's Metco Program." *American Economic Review* 94 (5):1613–1634.

Athey, Susan. 2019. "The Impact of Machine Learning on Economics." In *The economics of artificial intelligence*. University of Chicago Press, 507–552.

Athey, Susan and Guido W Imbens. 2019. "Machine learning methods that economists should know about." *Annual Review of Economics* 11:685–725.

Athey, Susan and Stefan Wager. 2019. "Estimating treatment effects with causal forests: An application." *arXiv preprint arXiv:1902.07409* .

Balestra, Simone, Beatrix Eugster, and Helge Liebert. 2020. "Summer-born struggle: The effect of school starting age on health, education, and work." *Health Economics* 29 (5):591–607.

———. forthcoming. "Peers with special needs: effects and policies." *Review of Economics and Statistics* .

Balestra, Simone, Aurélien Sallin, and Stefan Wolter. forthcoming. "High-ability influencers? The heterogeneous effects of gifted classmates." *Journal of Human Resources* .

Bertrand, Marianne and Jessica Pan. 2013. "The trouble with boys: social influences and the gender gap in disruptive behavior." *American Economic Journal: Applied Economics* 5 (1):32–64.

Bietenbeck, Jan. 2020. "The long-term impacts of low-achieving childhood peers: evidence from project STAR." *Journal of the European Economic Association* 18 (1):392–426.

Bifulco, Robert, Jason Fletcher, and Stephen 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.

Black, Sandra, Paul Devereaux, and Kjell Salvanes. 2013. "Under pressure? The effect of peers on outcomes of young adults." *Journal of Labor Economics* 31 (1):119–153.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. 2020. "Peer effects in networks: A survey." *Annual Review of Economics* 12:603–629.

Brenøe, Anne Ardila and Ulf Zölitz. 2020. "Exposure to more female peers widens the gender gap in stem participation." *Journal of Labor Economics* 38 (4):1009–1054.

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

Carrell, Scott and Mark Hoekstra. 2010. "Externalities in the classroom: how children exposed to domestic violence affect everyone's kids." *American Economic Journal: Applied Economics* 2 (1):211–228.

Carrell, Scott, Mark Hoekstra, and Elira Kuka. 2018. "The long-run effects of disruptive peers." *American Economic Review* 108 (11):3377–3415.

Eisenkopf, Gerald, Zohal Hessami, Urs Fischbacher, and Heinrich W Ursprung. 2015. "Academic performance and single-sex schooling: Evidence from a natural experiment in Switzerland." *Journal of economic behavior & organization* 115:123–143.

Figlio, David. 2007. "Boys named Sue: disruptive children and their peers." *Education Finance and Policy* 2 (4):376–394.

Fryer, Roland and Steven Levitt. 2010. "An empirical analysis of the gender gap in mathematics." *American Economic Journal: Applied Economics* 2 (2):210–240.

Graham, Bryan. 2011. "Econometric methods for the analysis of assignment problems in the presence of complementarity and social spillovers." In *Handbook of Social Economics*, vol. 1, edited by Jess Benhabib, Alberto Bisin, and Matthew Jackson. Elsevier, 965–1052.

Graham, Bryan, Guido Imbens, and Geert Ridder. 2010. "Measuring the effects of segregation in the presence of social spillovers: a nonparametric approach." Tech. rep., National Bureau of Economic Research.

Graham, Bryan, Geert Ridder, Petra M Thiemann, and Gema Zamarro. 2020. "Teacher-to-classroom assignment and student achievement." Tech. rep., National Bureau of Economic Research.

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, John Kain, and Steven Rivkin. 2009. "New evidence about Brown v. Board of Education: the complex effects of school racial composition on achievement." *Journal of Labor Economics* 27 (3):349–383.

Hofner, Benjamin, Luigi Boccuto, and Markus Göker. 2015. "Controlling false discoveries in high-dimensional situations: boosting with stability selection." *BMC bioinformatics* 16 (1):1–17.

Isphording, Ingo and Ulf Zölitz. 2020. "The value of a peer." Department of Economics Working Paper No. 342, University of Zurich.

Lavy, Victor, 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." *Economic Journal* 122 (559):208–237.

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

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

Lim, Michael and Trevor Hastie. 2015. "Learning interactions via hierarchical group-lasso regularization." *Journal of Computational and Graphical Statistics* 24 (3):627–654.

Manski, Charles. 1993. "Identification of endogenous social effects: the reflection problem." *Review of Economic Studies* 60 (3):531–542.

Meinshausen, Nicolai and Peter Bühlmann. 2010. "Stability selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (4):417–473.

Mullainathan, Sendhil and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2):87–106.

Niederle, Muriel and Lise Vesterlund. 2010. "Explaining the gender gap in math test scores: The role of competition." *Journal of Economic Perspectives* 24 (2):129–44.

Pregaldini, Damiano, Uschi Backes-Gellner, and Gerald Eisenkopf. 2020. "Girls' preferences for STEM and the effects of classroom gender composition: New evidence from a natural experiment." *Journal of Economic Behavior & Organization* 178:102–123.

Sacerdote, Bruce. 2014. "Experimental and quasi-experimental analysis of peer effects: two steps forward?" *Annual Review of Economics* 6 (1):253–272.

Sallin, Aurélien. 2021. "Estimating returns to special education: combining machine learning and text analysis to address confounding." *arXiv preprint arXiv:2110.08807* .

Shah, Rajen and Richard Samworth. 2013. "Variable selection with error control: another look at stability selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (1):55–80.

Swiss Federal Statistical Office. 2020. "Statistik der Sonderpädagogik – Schuljahr 2018/19." Annual Report, Federal Department of Home Affairs.

Yuan, Ming and Yi Lin. 2006. "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (1):49–67.

# Tables and Figures

| | Mean | Std.dev. | Min | Max | *N* |
|---|---|---|---|---|---|
| **A: Individual binary characteristics (types)** | | | | | |
| Older age | 0.17 | 0.38 | | | 48,714 |
| Identified with special needs | 0.29 | 0.45 | | | 48,714 |
| Gifted | 0.01 | 0.10 | | | 48,714 |
| Nonnative | 0.15 | 0.36 | | | 48,714 |
| Female | 0.50 | 0.50 | | | 48,714 |
| **B: Cohort peers within schools (leave-own-out)** | | | | | |
| Older peers | 0.172 | 0.113 | 0.00 | 0.94 | 48,714 |
| Peers with special needs | 0.291 | 0.180 | 0.00 | 1.00 | 48,714 |
| Gifted peers | 0.010 | 0.021 | 0.00 | 0.23 | 48,714 |
| Nonnative peers | 0.149 | 0.183 | 0.00 | 0.94 | 48,714 |
| Female peers | 0.498 | 0.158 | 0.00 | 1.00 | 48,714 |
| **C: Classroom peers within school-track-years (leave-own-out)** | | | | | |
| Older peers | 0.172 | 0.134 | 0.00 | 1.00 | 48,714 |
| Peers with special needs | 0.291 | 0.197 | 0.00 | 1.00 | 48,714 |
| Gifted peers | 0.010 | 0.027 | 0.00 | 0.40 | 48,714 |
| Nonnative peers | 0.149 | 0.201 | 0.00 | 1.00 | 48,714 |
| Female peers | 0.498 | 0.169 | 0.00 | 1.00 | 48,714 |
| **C: Cells** | | | | | |
| Year of test | 2012 | 2.86 | 2008 | 2017 | 48,714 |
| Class size | 19.13 | 3.63 | 10 | 30 | 48,714 |
| Cohort size | 46.44 | 22.77 | 11 | 118 | 48,714 |
| **D: Outcomes** | | | | | |
| Composite test score | 0.00 | 1.00 | -4.66 | 4.2 | 48,714 |

Summary statistics for the population of students in the inclusive school system of the Canton of St. Gallen. Information for the cohorts and the classroom composition is given. Number of classrooms: 2,674; Number of school-track-years: 1,357; Number of school-tracks: 142. *Source: SPS*

Table 1: Summary statistics

| | Observed | | | | Randomized | | Difference |
|---|---|---|---|---|---|---|---|
| | Mean | Std.dev. | Min | Max | Mean | Std.dev. | P.value |
| **A: Cohorts within school** | | | | | | | |
| **Raw cohort variables** | | | | | | | |
| Older peers | 0.172 | 0.113 | 0 | 0.94 | 0.187 | 0.117 | 0.554 |
| Peers with special needs | 0.290 | 0.180 | 0 | 1.00 | 0.315 | 0.184 | 0.919 |
| Gifted peers | 0.010 | 0.021 | 0 | 0.23 | 0.010 | 0.021 | 0.616 |
| Nonnative peers | 0.149 | 0.183 | 0 | 0.94 | 0.156 | 0.163 | 0.922 |
| Female peers | 0.498 | 0.158 | 0 | 1.00 | 0.489 | 0.146 | 0.983 |
| **B: Cohorts within school** | | | | | | | |
| **Residuals after removing school-track fixed effects and year trends** | | | | | | | |
| Older peers | 0.00 | 0.069 | -0.32 | 0.35 | 0.00 | 0.072 | 0.258 |
| Peers with special needs | 0.00 | 0.079 | -0.32 | 0.35 | 0.00 | 0.082 | 0.787 |
| Gifted peers | 0.00 | 0.016 | -0.07 | 0.18 | 0.00 | 0.017 | 0.479 |
| Nonnative peers | 0.00 | 0.110 | -0.60 | 0.62 | 0.00 | 0.063 | 0.836 |
| Female peers | 0.00 | 0.083 | -0.35 | 0.44 | 0.00 | 0.091 | 0.996 |
| **C: Classrooms within school-track-years** | | | | | | | |
| **Raw classroom variables** | | | | | | | |
| Older peers | 0.172 | 0.134 | 0 | 0.95 | 0.178 | 0.132 | 0.803 |
| Peers with special needs | 0.290 | 0.297 | 0 | 1.00 | 0.305 | 0.200 | 0.783 |
| Gifted peers | 0.010 | 0.027 | 0 | 0.36 | 0.010 | 0.027 | 0.990 |
| Nonnative peers | 0.149 | 0.201 | 0 | 1.00 | 0.158 | 0.200 | 0.890 |
| Female peers | 0.498 | 0.169 | 0 | 1.00 | 0.492 | 0.173 | 0.915 |
| **D: Classrooms within school-track-years** | | | | | | | |
| **Residuals after removing school-track-year fixed effects** | | | | | | | |
| Older peers | 0.00 | 0.072 | -0.49 | 0.82 | 0.00 | 0.066 | 0.528 |
| Peers with special needs | 0.00 | 0.082 | -0.62 | 0.80 | 0.00 | 0.076 | 0.366 |
| Gifted peers | 0.00 | 0.020 | -0.12 | 0.17 | 0.00 | 0.017 | 0.966 |
| Nonnative peers | 0.00 | 0.084 | -0.55 | 0.71 | 0.00 | 0.058 | 0.649 |
| Female peers | 0.00 | 0.063 | -0.57 | 0.62 | 0.00 | 0.086 | 0.670 |

*Notes:* Variation in cohort or classroom composition measures after removing school-track fixed effects (and time trends or school-track-year fixed effects.) Randomization checks with 500 draws. For each random draw, students are randomly reassigned to classes or cohorts within the same school-tracks or school-track-years. The presented standard deviation is the mean standard deviations across the 500 reassignments. The test for mean differences between the random draw and the observed data is a $t$-test, and p-values are reported.

Table 2: Variation in cohort or classroom composition measures and randomization checks

| | Mean | Std.dev. |
|---|---|---|
| **A: Cohorts with highest average test score** | | |
| Older peers | -0.064*** | 0.007 |
| Peers with special needs | -0.043*** | 0.010 |
| Gifted peers | 0.007*** | 0.002 |
| Nonnative peers | -0.045*** | 0.007 |
| Female peers | -0.092*** | 0.015 |
| Cohort size | -6.348*** | 1.403 |
| **B: Classrooms with highest average test score** | | |
| Older peers | -0.103*** | 0.006 |
| Peers with special needs | -0.139*** | 0.003 |
| Gifted peers | 0.019*** | 0.005 |
| Nonnative peers | -0.175*** | 0.009 |
| Female peers | -0.055*** | 0.011 |
| Classroom size | -0.758*** | 1.078 |

$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.001$

*Notes:* This table shows the composition of the cohorts and classrooms that maximize average test scores (in panels A and C).All effects are demeaned at the level of randomization (school-track for cohorts, school-track-years for classrooms). The interpretation of mean coefficients is relative to the mean at the level of randomization. Bootstrapped standard errors are reported.

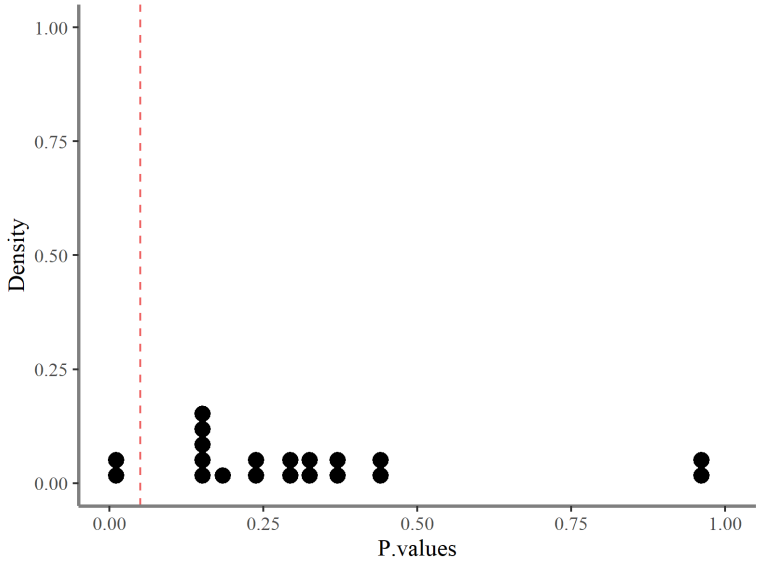
Table 3: Optimal classroom and cohort environments

| | Randomized regime $\frac{1}{N}\sum_{i=1}^{N}\left[\hat{Y}_{ic}^{\text{random}}\right]$ | Manipulated regime $\frac{1}{N}\sum_{i=1}^{N}\left[\hat{Y}_{ic}^{\text{segr.}}\right]$ | Difference *ACE* |
|---|---|---|---|
| **A: Variation in aggregated test score** | | | |
| **Removed types**: | | | |
| Older peers | 0.012 | 0.010 | -0.003*** |
| Peers with special needs | 0.010 | -0.005 | -0.015*** |
| Nonnative peers | 0.012 | 0.017 | 0.005*** |
| Female peers | 0.010 | 0.009 | -0.001** |
| **B: Corresponding Gini coefficients** | | | |
| **Removed types**: | | | |
| Older peers | 0.222 | 0.210 | -0.012*** |
| Peers with special needs | 0.225 | 0.214 | -0.011*** |
| Nonnative peers | 0.226 | 0.223 | -0.003*** |
| Female peers | 0.227 | 0.223 | -0.004 *** |

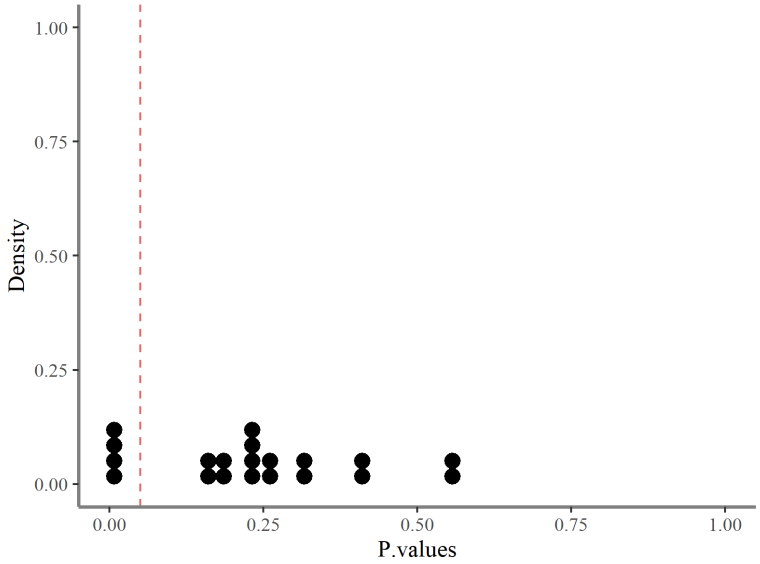
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.001$

*Notes:* This table shows the predicted school average test score under both the random and manipulated (simulated) allocation regimes. The "random" classroom allocation scheme has random allocation of all students to classrooms, whereas the "manipulated" classroom allocation scheme removes one student with the type of interest from all classrooms but one, and creates a cluster of students with the type of interest in the left-out classroom. The manipulation dimensions are the main types used in the paper (except gifted students, as the category is very small). The difference shows the *average counterfactual effect (ACE)*. All effects are demeaned at the school-track-years level. For each aggregated test score comparison, Panel B shows the variation in the Gini coefficient. For each simulation, 500 random draws are conducted, and reported standard errors are bootstrapped.

Table 4: Comparison of randomized and counterfactual manipulated allocation




(a) Balancing check for cohort identification




(b) Balancing check for classroom identification

Figure 1: Balancing checks for cohort and classroom identification

*Notes:* This graph shows the distribution of p-values obtained from regressing each particular type on the proportion of peers of all other types in the relevant cells (cohort or classroom). Five types and their five corresponding proportions are used, giving $5 \times 5 - 5 = 20$ regressions. The regressions of type share on their corresponding own type are excluded. P-values are the p.values of the type coefficient of each regression. Each regression controls for the relevant fixed effects (school-track.year for classroom identification and school-track for cohort identification), as well as for the mean peer type at the level of randomization (this last control is the "correction" proposed by Guryan, Kroft, and Notowidigdo 2009).

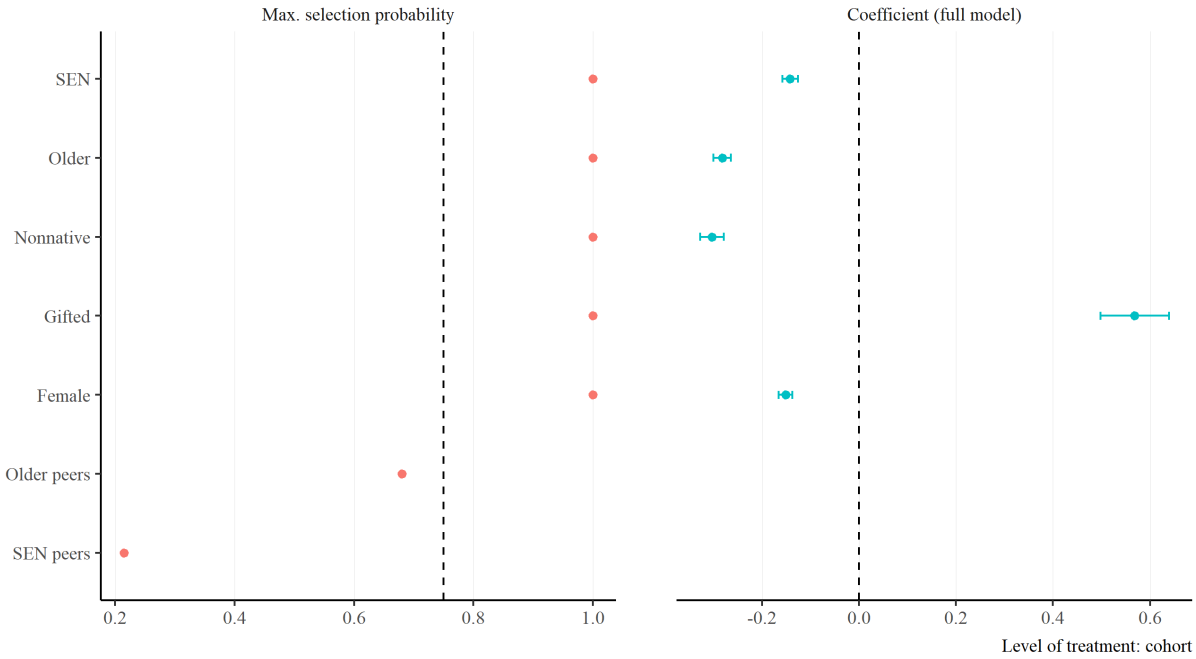



Figure 2: Stability selection and effect size at the cohort level without interactions

*Notes:* the left panel of this graph reports the probability for a variable to be selected. The learning algorithm is stable selection with hierarchical group lasso on 200 folds of size $n/2$. Selected variables are variable selected with a probability higher than 0.75 (dashed line). The $\times$ indicates interactions, the term "peers" indicate peer effects, and the term "size" is the size of the cell (either classroom or cohort). The right panel displays the OLS coefficients and their 95% confidence intervals only for the variable selected. More details on how coefficients are computed can be found in Section 3.3. Effect sizes and confidence intervals are presented for information only and must be interpreted with caution.

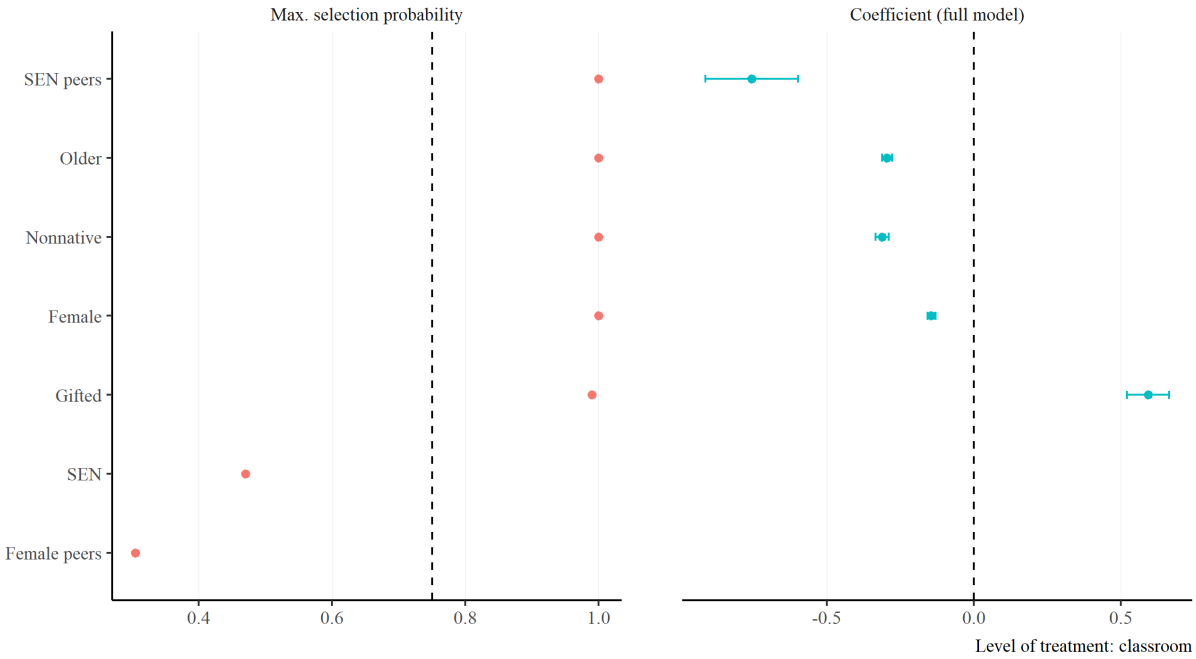



Figure 3: Stability selection and effect size at the classroom level without interactions

*Notes:* the left panel of this graph reports the probability for a variable to be selected. The learning algorithm is stable selection with hierarchical group lasso on 200 folds of size $n/2$. Selected variables are variable selected with a probability higher than 0.75 (dashed line). The $\times$ indicates interactions, the term "peers" indicate peer effects, and the term "size" is the size of the cell (either classroom or cohort). The right panel displays the OLS coefficients and their 95% confidence intervals only for the variable selected. More details on how coefficients are computed can be found in Section 3.3. Effect sizes and confidence intervals are presented for information only and must be interpreted with caution.

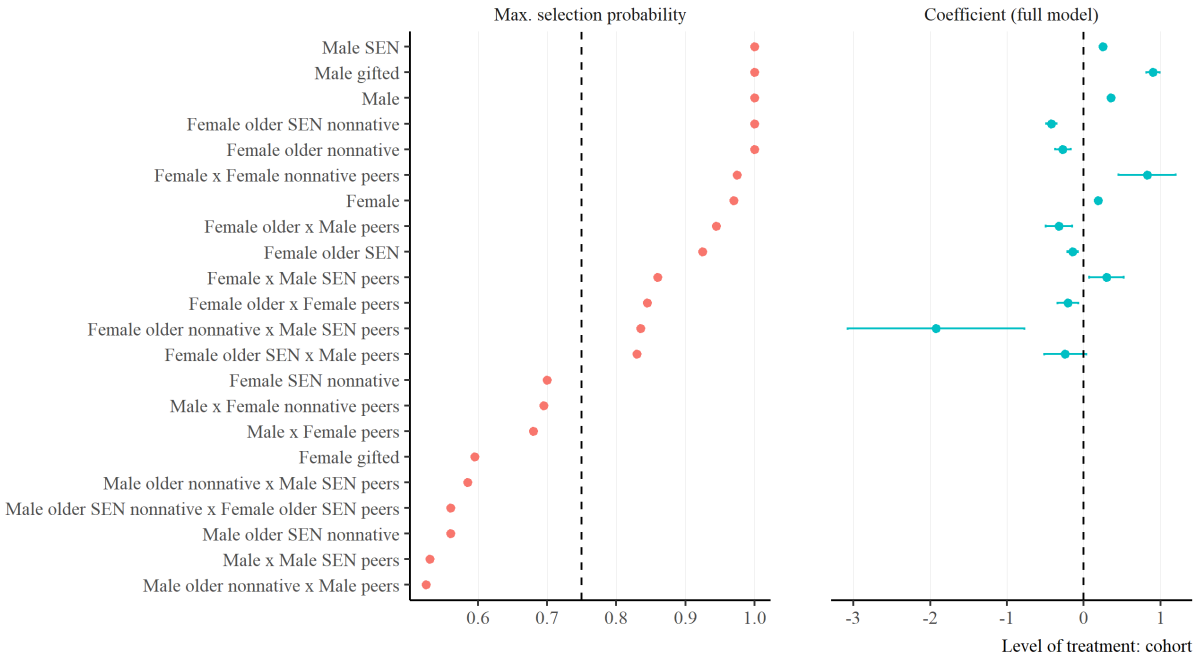



Figure 4: Stability selection and effect size at the cohort level with fully interacted types

*Notes:* the left panel of this graph reports the probability for a variable to be selected. The learning algorithm is stable selection with hierarchical group lasso on 200 folds of size $n/2$. Selected variables are variable selected with a probability higher than 0.75 (dashed line). The $\times$ indicates interactions, the term "peers" indicate peer effects, and the term "size" is the size of the cell (either classroom or cohort). The right panel displays the OLS coefficients and their 95% confidence intervals only for the variable selected. More details on how coefficients are computed can be found in Section 3.3. Effect sizes and confidence intervals are presented for information only and must be interpreted with caution.

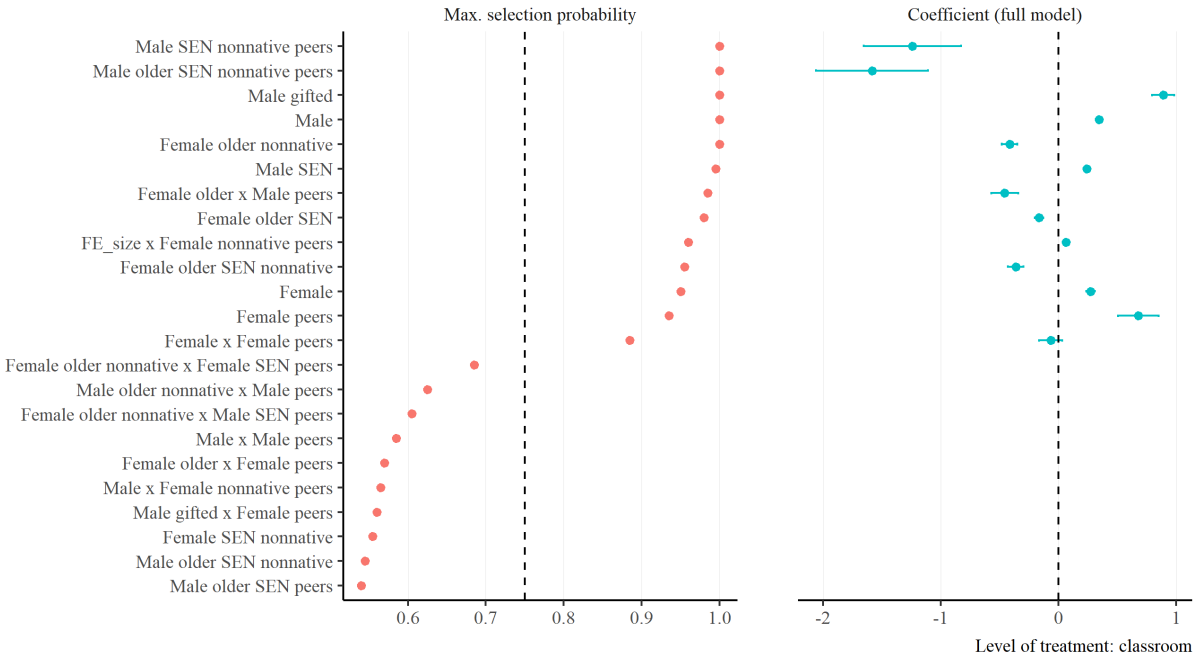



Figure 5: Stability selection and effect size at the classroom level with fully interacted types

*Notes:* the left panel of this graph reports the probability for a variable to be selected. The learning algorithm is stable selection with hierarchical group lasso on 200 folds of size $n/2$. Selected variables are variable selected with a probability higher than 0.75 (dashed line). The $\times$ indicates interactions, the term "peers" indicate peer effects, and the term "size" is the size of the cell (either classroom or cohort). The right panel displays the OLS coefficients and their 95% confidence intervals only for the variable selected. More details on how coefficients are computed can be found in Section 3.3. Effect sizes and confidence intervals are presented for information only and must be interpreted with caution.

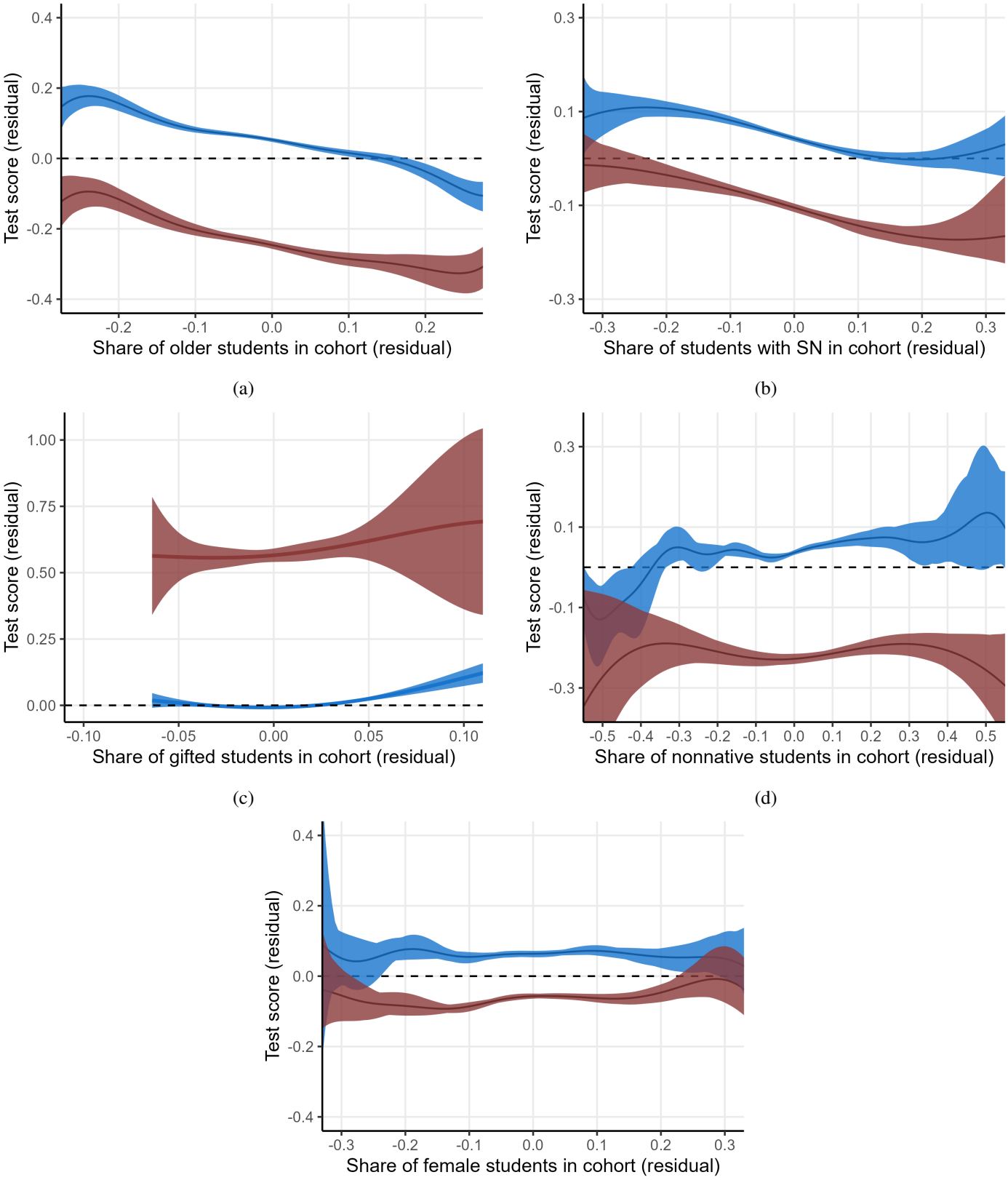
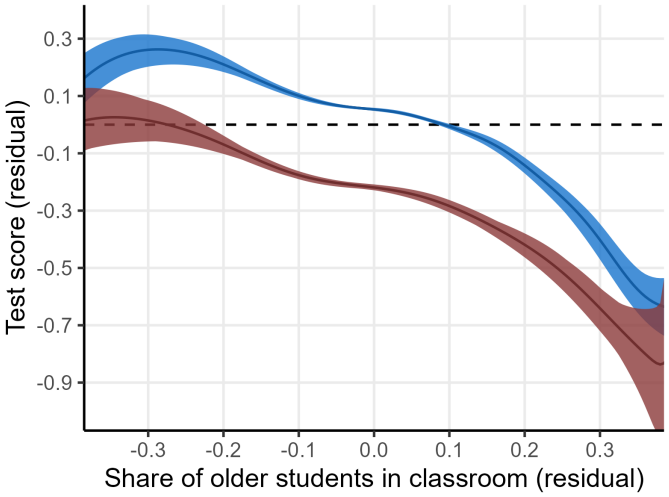
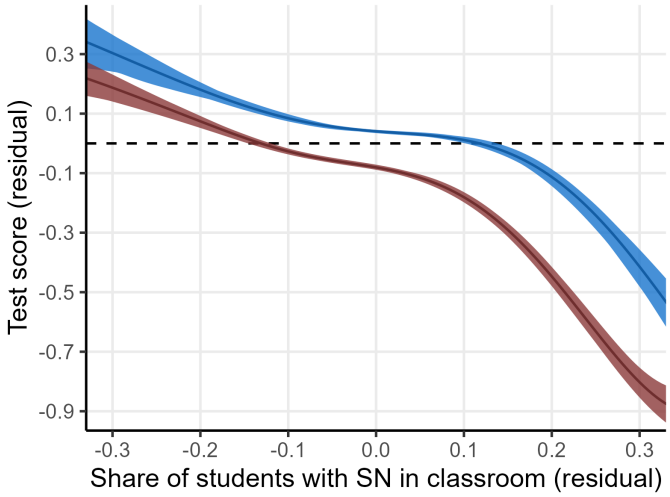



Figure 6: High-dimensional peer effects at the cohort level

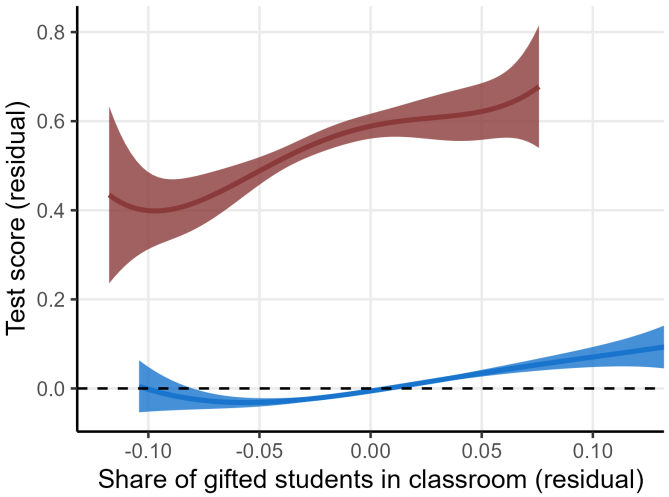
*Notes:* This figure displays the effect of the share of types of interest per cohort on the standardized test scores. Test scores and shares are demeaned from their school-track averages. Heterogeneity across own types is depicted in different colors: the own effect is given in red, and the effect for the other category is given in blue. 95% confidence intervals are represented from 50 predictions obtained with clustered cross-validation; at each point, the maximal and minimal upper and lower confidence intervals across the 50 predictions are represented.



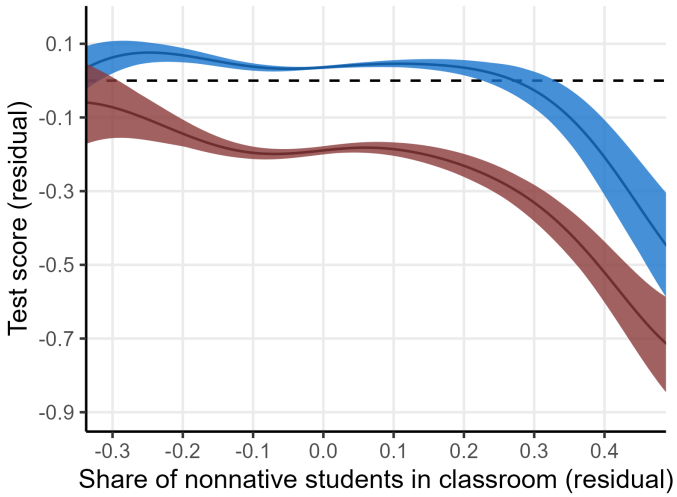

(a)



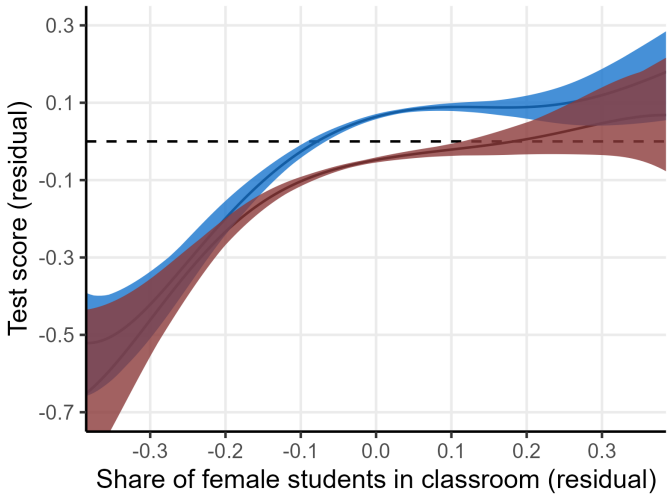

(b)




(c)




(d)




(e)

Figure 7: High-dimensional peer effects at the classroom level

*Notes:* This figure displays the effect of the share of types of interest per classroom on the standardized test scores. Test scores and shares are demeaned from their school-track-year averages. Heterogeneity across own types is depicted in different colors: the own effect is given in red, and the effect for the other category is given in blue. 95% confidence intervals are represented from 50 predictions obtained with clustered cross-validation; at each point, the maximal and minimal upper and lower intervals across the 50 predictions are represented.

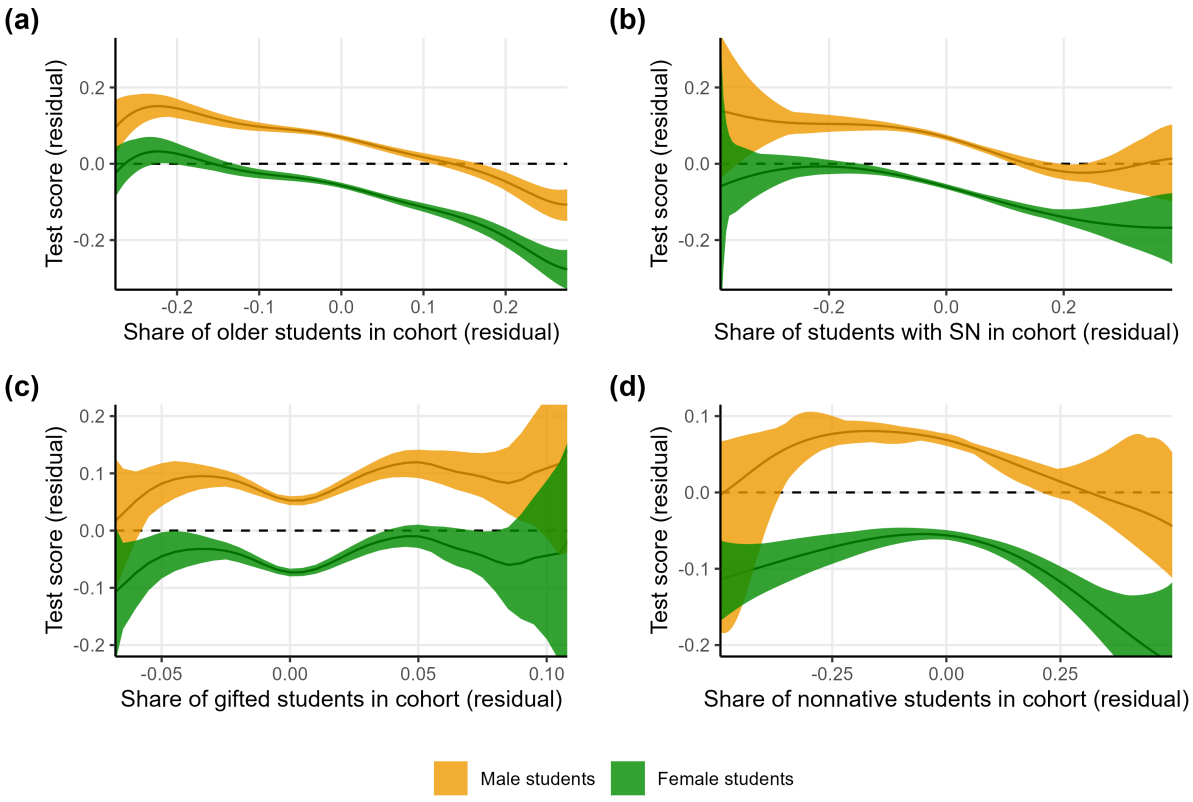



Figure 8: High-dimensional peer effects: gender heterogeneity at the cohort level

*Notes:* This figure displays the effect of the share of types of interest per cohort on the standardized test scores. Test scores and shares are demeaned from their school-track averages. Heterogeneity across gender is depicted in different colors for both genders. 95% confidence intervals are represented from 50 predictions obtained with clustered cross-validation; at each point, the maximal and minimal upper and lower intervals across the 50 predictions are represented.

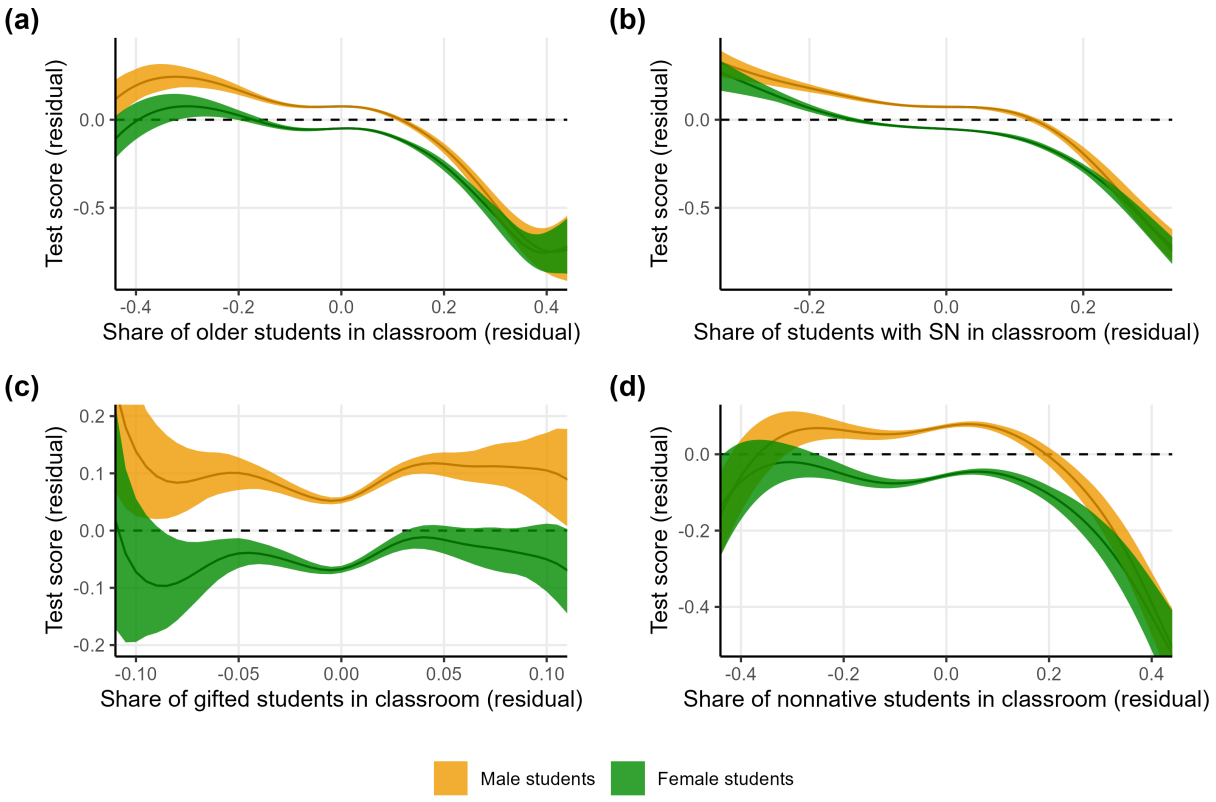
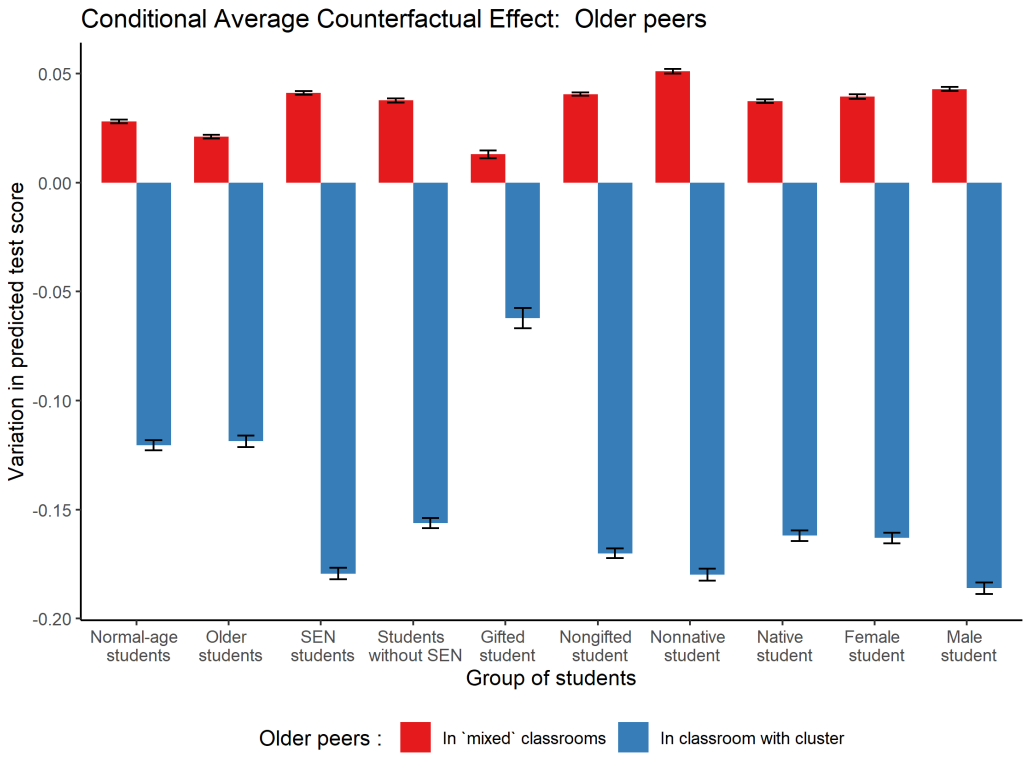
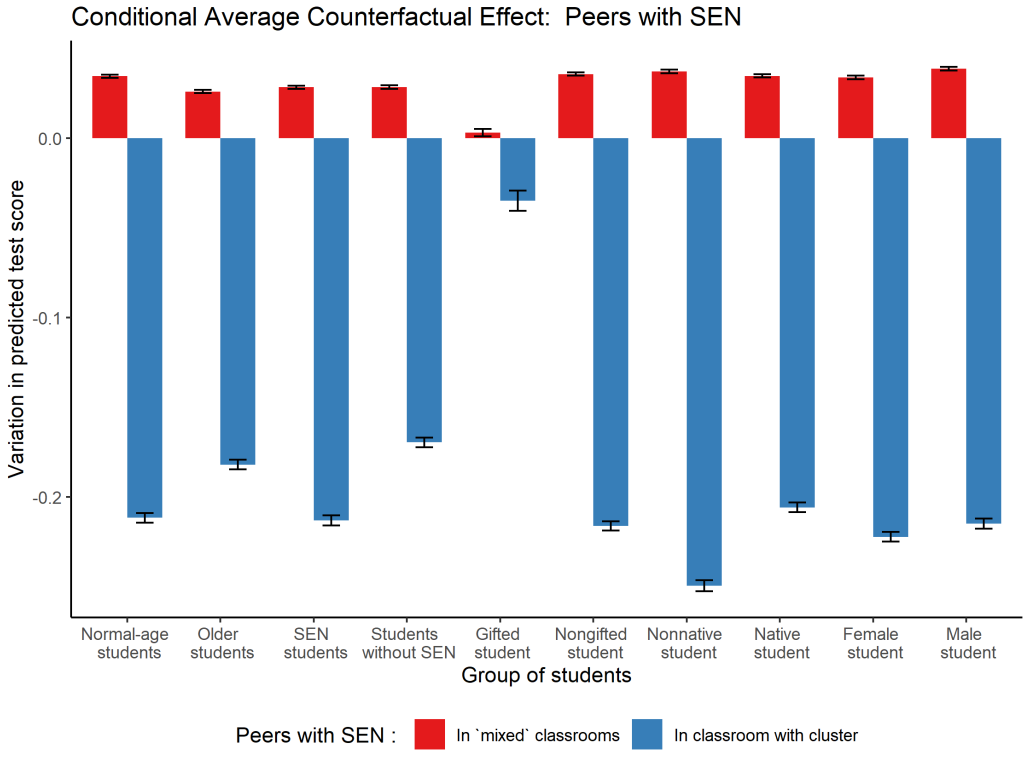



Figure 9: High-dimensional peer effects: gender heterogeneity at the classroom level

*Notes:* This figure displays the effect of the share of types of interest per classroom on the standardized test scores. Test scores and shares are demeaned from their school-track-year averages. Heterogeneity across gender is depicted in different colors for both genders. 95% confidence intervals are represented from 50 predictions obtained with clustered cross-validation; at each point, the maximal and minimal upper and lower intervals across the 50 predictions are represented.



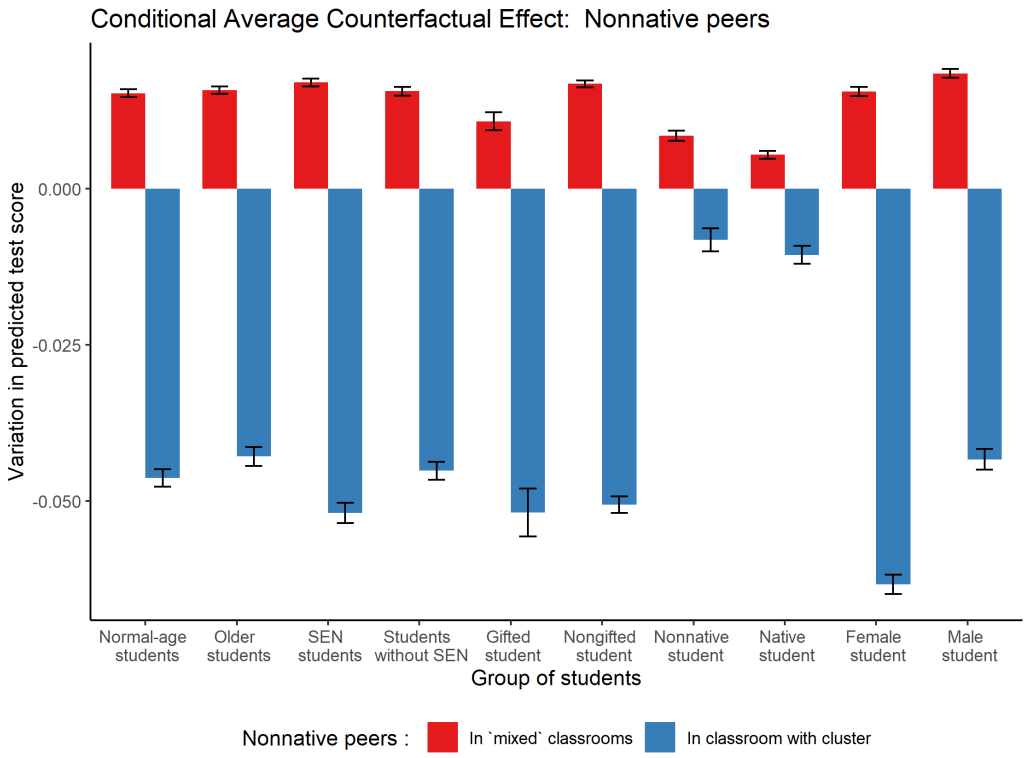

(a)



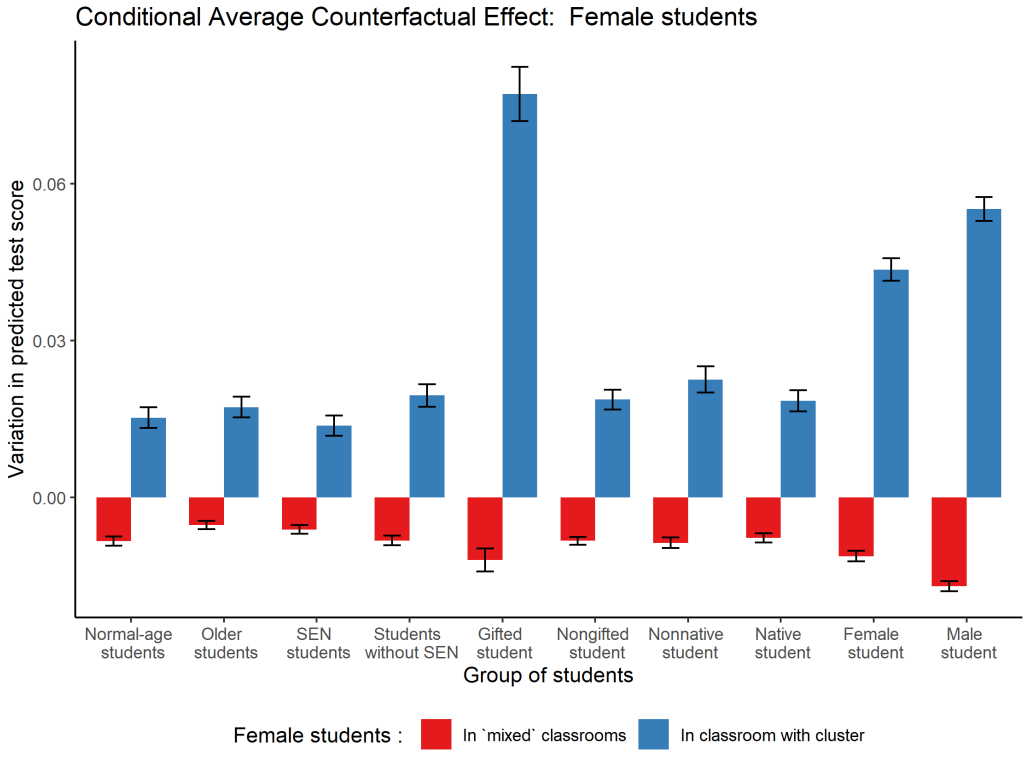

(b)

Figure 10: CACE: clustering of older peers and peers with SEN

*Notes:* This figure displays the *conditional average counterfactual effects (CACE)* of classroom allocation manipulation for the types of older and SEN. Each bar represent the difference in group test score averages under the different counterfactual regimes for each particular type of students. The color indicates the classroom: "in mixed classrooms" is the *CACE* for students kept in the mixed classrooms, whereas "in classroom with cluster" is the *CACE* for students who are in the left-out classroom with the cluster of marginally segregated students. Confidence intervals of 95% are obtained by bootstrapping (see main text).




(a)




(b)

Figure 11: CACE: clustering of nonnative and female students

*Notes:* This figure displays the *conditional average counterfactual effects (CACE)* of classroom allocation manipulation for the types of nonnative and female. Each bar represent the difference in group test score averages under the different counterfactual regimes for each particular type of students. The color indicates the classroom: "in mixed classrooms" is the *CACE* for students kept in the mixed classrooms, whereas "in classroom with cluster" is the *CACE* for students who are in the left-out classroom with the cluster of marginally segregated students. Confidence intervals of 95% are obtained by bootstrapping (see main text).

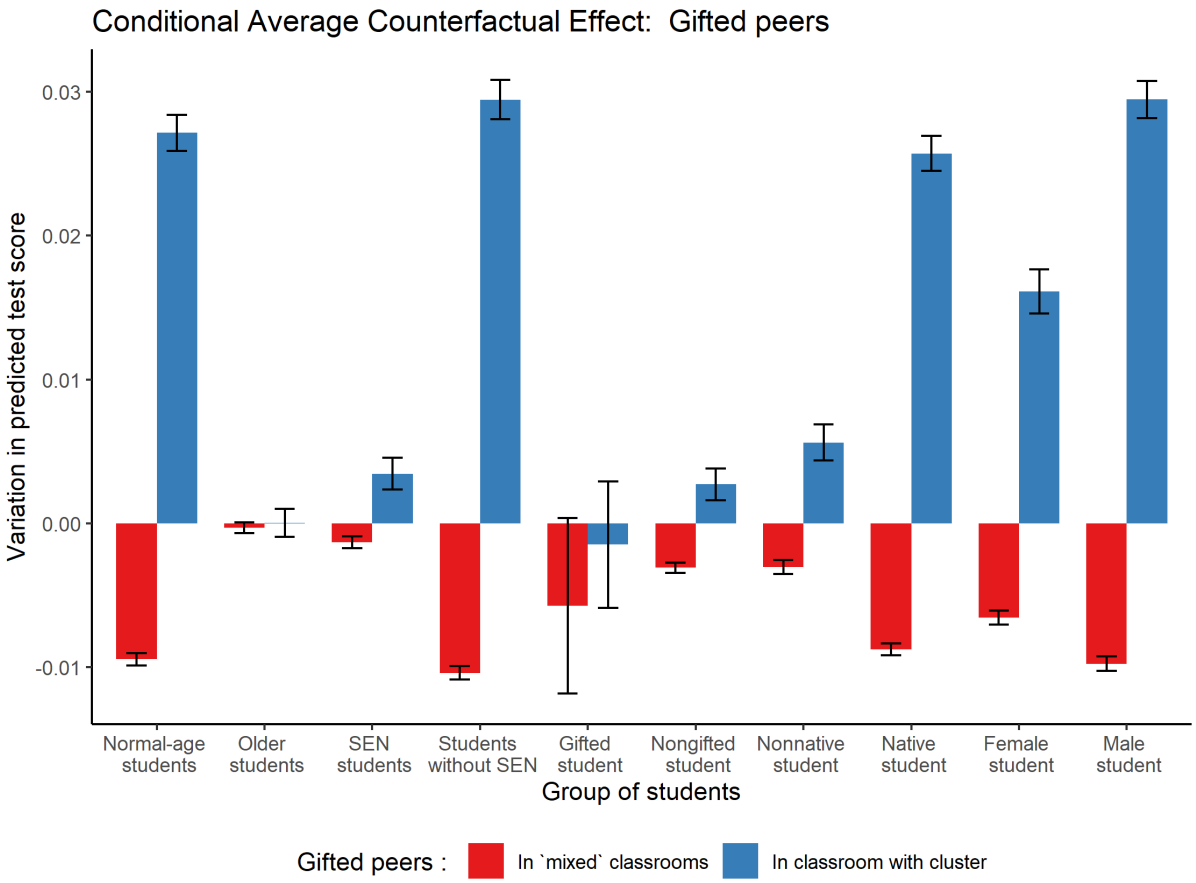



Figure 12: CACE: clustering of gifted students

*Notes:* This figure displays the *conditional average counterfactual effects (CACE)* of segregation for the gifted type. Each bar represent the difference in group test score averages under the different counterfactual regimes for each particular type of students. The color indicates the classroom: "in mixed classrooms" is the *CACE* for students kept in the mixed classrooms, whereas "in classroom with cluster" is the *CACE* for students who are in the left-out classroom with the cluster of marginally segregated gifted students. Confidence intervals of 95% are obtained by bootstrapping (see main text).

# Appendix

## A Supplementary Material

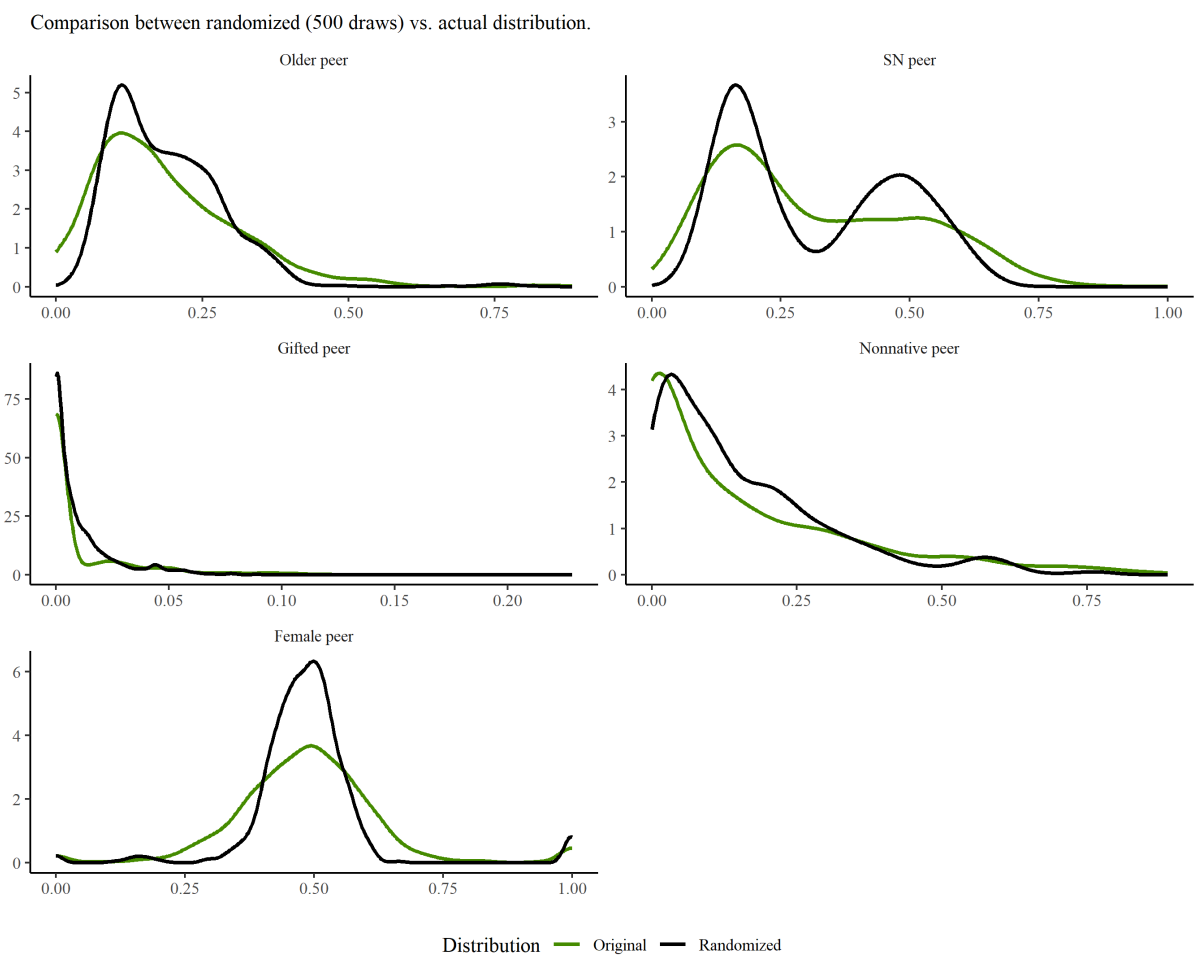



Figure A.1: Balancing check for cohort identification

*Notes:* This graph shows the actual and the randomized (simulated) distribution of the proportion of peers of a given type within school-tracks. An observation is a cohort. We randomly draw observations at the school-track level 500 times with replacement and compare the randomly sampled distribution of student types with the distribution we observe in the data.

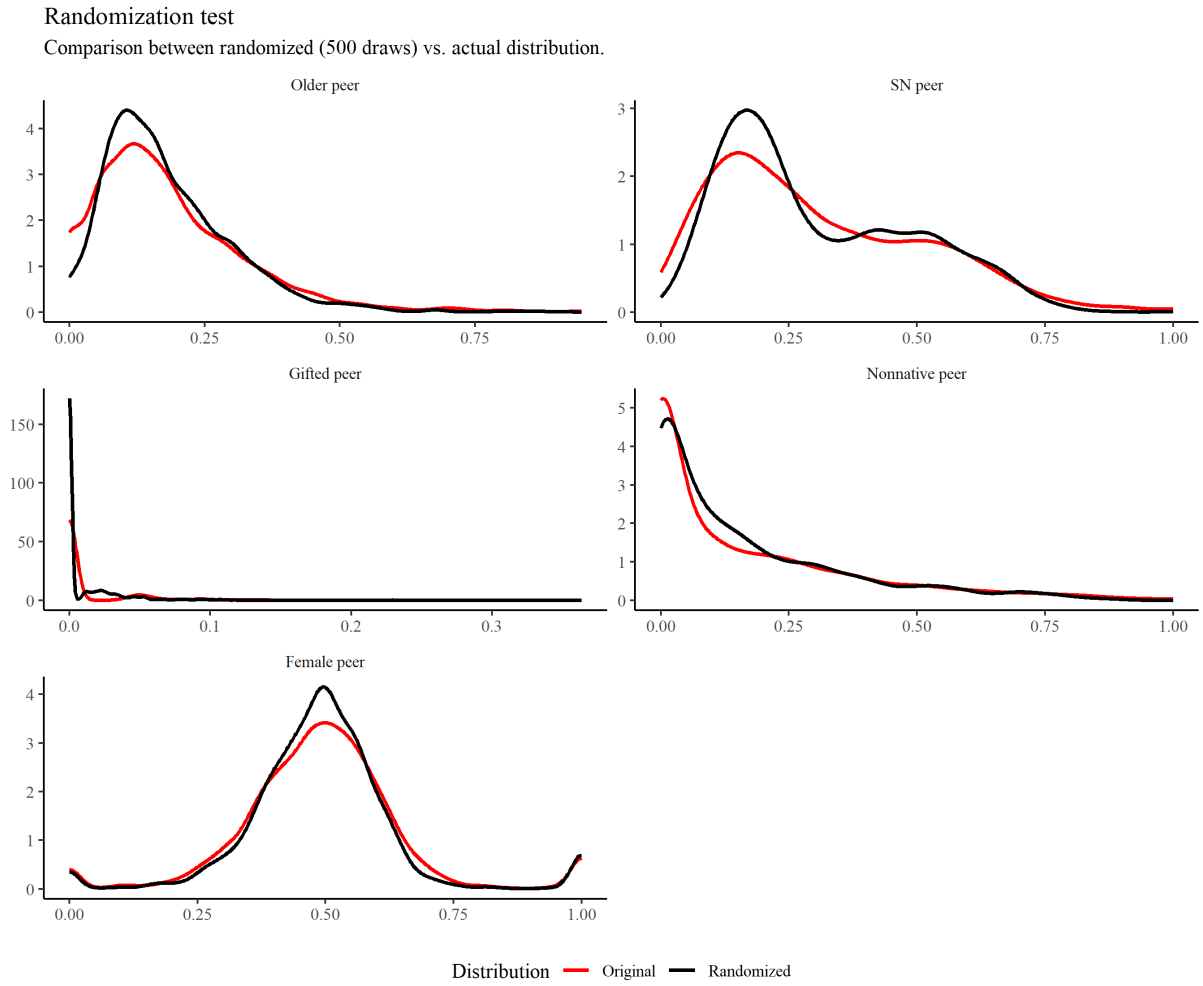



Figure A.2: Balancing check for classroom identification

*Notes:* This graph shows the actual and the randomized (simulated) distribution of the proportion of peers of a given type within school-track-years. An observation is a classroom. We randomly draw observations at the school-track-year level 500 times with replacement and compare the randomly sampled distribution of student types with the distribution we observe in the data.

## B Appendix: Stable selection with interactions of degree 2

As presented in Figures B.1 and B.2, 13 variables are selected at the cohort level, and 19 at the classroom level. Among the variables selected at the cohort level, only four variables are peer effects: the effect of older peers, and the effect of peers with SEN. These two peer effects are heterogeneous: older peers have different effects on nonnative and native students, and the effects of older peers and of peers with SEN have an interacted effect, meaning that the effect of older peers changes as a function of the proportion of peers with SEN in the cohort. This is a case of strong hierarchy: the interaction effects are selected together with their main effects. The other variables selected are all types or interactions between types, which is a first indication that our five main types hide substantial heterogeneity.

At the classroom level, the five main types are selected (although the SEN status is only marginally selected). The first dominating peer effects are spillovers from peers with SEN. The algorithm selects, in 100% of cases, peer effects from students with SEN on other students with SEN, and peer effects from students with SEN on nonnatives. However, and surprisingly, the main effect of peers with SEN is not selected: this interesting case of weak hierarchy means that the main effect of classmates with SEN is not part of the "true" model. Thus, the stable selection algorithm gives us a more refined understanding of peer effects from students with SEN, who affect mostly their nonnative classmates and their classmates with SEN. The second dominating effects in the classroom are effects from older peers (see also Bietenbeck, 2020): older peers have a negative impact on their peers, and this impact is interacted with the classroom size and with the effect of female peers. Finally, some variables are interacted with the classroom size: the effect of peers with SEN, the effect of older peers, and the own SEN status. The influence of classroom size for students who are more likely to fall behind is anything but surprising: their academic success is more likely to depend on the availability of teaching resources and individual teacher attention.

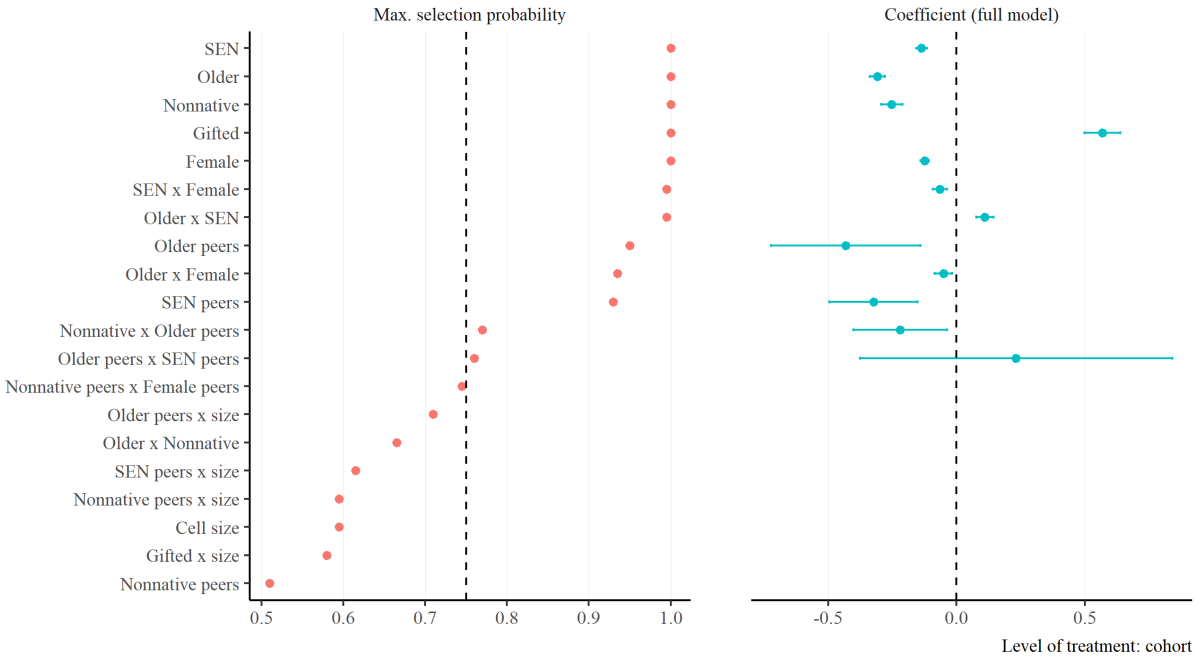



Figure B.1: Stability selection and effect size at the cohort level with interactions of level 2

*Notes:* the left panel of this graph reports the probability for a variable to be selected. The learning algorithm is stable selection with hierarchical group lasso on 200 folds of size $n/2$. Selected variables are variable selected with a probability higher than 0.75 (dashed line). The $\times$ indicates interactions, the term "peers" indicate peer effects, and the term "size" is the size of the cell (either classroom or cohort). The right panel displays the OLS coefficients and their 95% confidence intervals only for the variable selected. More details on how coefficients are computed can be found in Section 3.3. Effect sizes and confidence intervals are presented for information only and must be interpreted with caution.

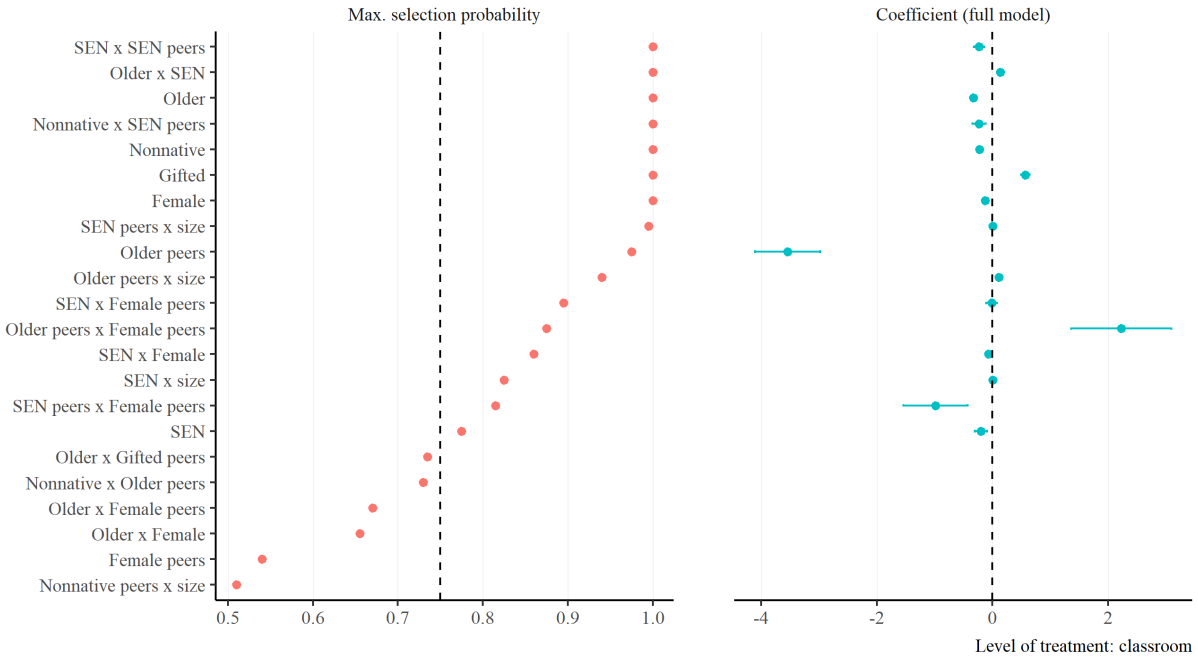



Figure B.2: Stability selection and effect size at the classroom level with interactions of level 2

*Notes:* the left panel of this graph reports the probability for a variable to be selected. The learning algorithm is stable selection with hierarchical group lasso on 200 folds of size $n/2$. Selected variables are variable selected with a probability higher than 0.75 (dashed line). The $\times$ indicates interactions, the term "peers" indicate peer effects, and the term "size" is the size of the cell (either classroom or cohort). The right panel displays the OLS coefficients and their 95% confidence intervals only for the variable selected. More details on how coefficients are computed can be found in Section 3.3. Effect sizes and confidence intervals are presented for information only and must be interpreted with caution.

## C Appendix: Counterfactual analysis with full segregation

In this part, we are interested in the *ACE* and *CACE* of *full* segregation. We proceed as follows: we randomly draw a pool of 100 students out of the main sample (our "school"), and we randomly create 5 classrooms of 20 students each. In a second step, we create counterfactual classrooms: we put all students with a particular type (e.g., all students with SEN) in segregated classrooms, and we assign the other students randomly to the other classrooms. If a segregated classroom is not filled, we fill the rest of the classroom seats to random students. For instance, if there are 30 students with SEN in our school, we fill the first segregated classroom entirely with students with SEN, and the second classroom is filled with 10 students with SEN and 10 students without SEN. For each student and for both settings, we generate the LoO at the classroom level along the five main characteristics. To obtain the individual predicted values, we match our randomly drawn observations with their nearest neighbors in the original sample. We do this exercise for 100 random drawn, and for each of the $M$ predicted values. We compute bootstrapped confidence intervals (as we have, for each random sample draw, $M$ predicted values).

**Results for the *ACE*** Results for the *ACE* are presented in Table C.1. The picture is clear: segregating older peers and peers with SEN has a clear negative impact on the overall aggregated academic performance. However, we find that the segregation of nonnative students as well as the segregation of female students do not have any aggregate welfare consequences. All segregated settings slightly reduce the Gini coefficient, which means that, as expected, segregation increases inequality.

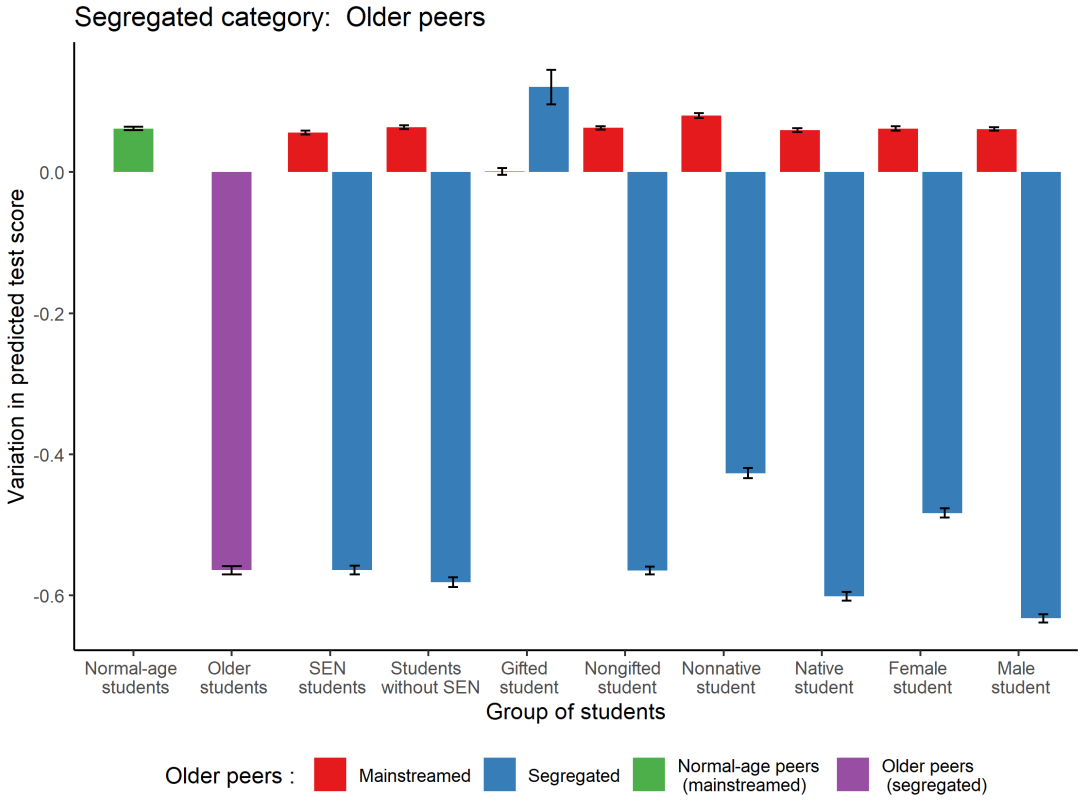
**Results for the *CACE*** We look at the variations in predicted aggregate test scores for each type of student. Figure C.1 and Figure C.2 present the gains and losses for students of all categories when older students (Figure C.1a.), when students with SEN (Figure C.1b.), when nonnative students (Figure C.2a.), and when female students (Figure C.2b.) are segregated. These are group test score averages under the different counterfactual regimes.

| | Randomized regime $\frac{1}{N}\sum_{i=1}^{N}\left[\hat{Y}_{ic}^{\text{random}}\right]$ | Segregated regime $\frac{1}{N}\sum_{i=1}^{N}\left[\hat{Y}_{ic}^{\text{segr.}}\right]$ | Difference *ACE* |
|---|---|---|---|
| **A: Variation in aggregated test score** | | | |
| **Segregation dimension**: | | | |
| Older peers | 0.014 | -0.031 | -0.045*** |
| Peers with special needs | 0.010 | -0.103 | -0.114*** |
| Nonnative peers | 0.011 | 0.002 | -0.001*** |
| Female peers | 0.016 | 0.010 | -0.006 *** |
| **B: Corresponding Gini coefficients** | | | |
| **Segregation dimension**: | | | |
| Older peers | 0.230 | 0.213 | -0.017*** |
| Peers with special needs | 0.223 | 0.265 | 0.041*** |
| Nonnative peers | 0.225 | 0.190 | -0.036*** |
| Female peers | 0.230 | 0.187 | -0.043 *** |

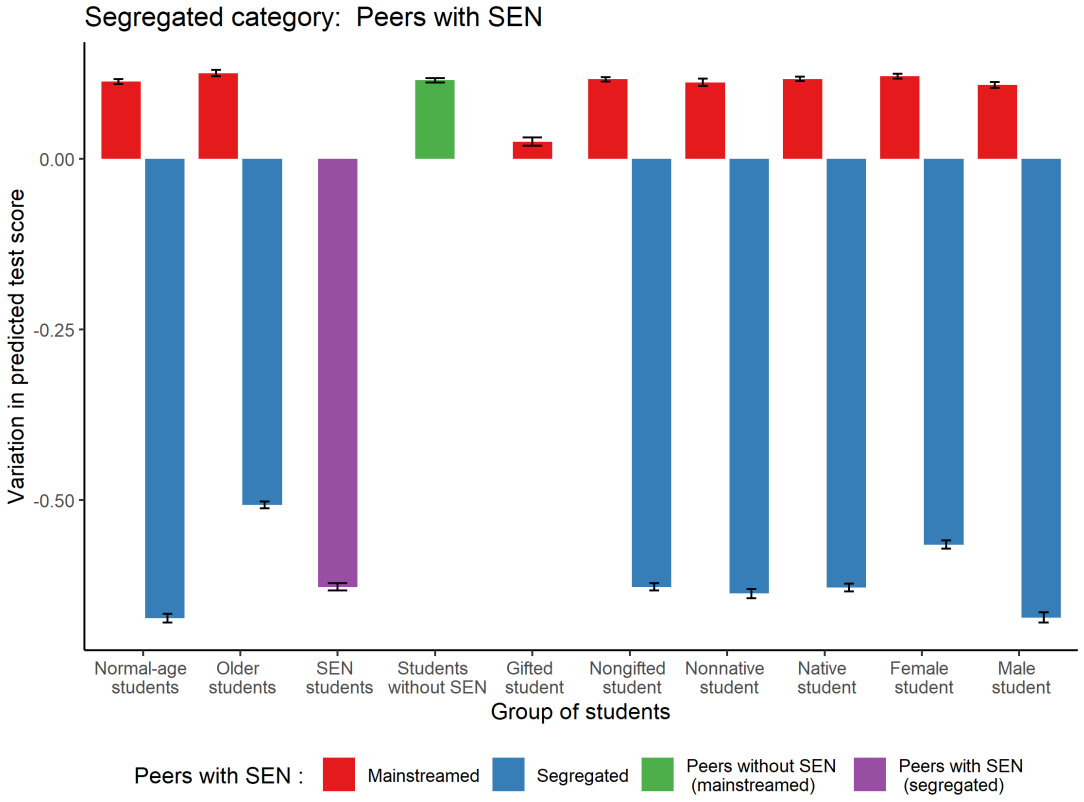
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.001$

Table C.1: Comparison of randomized and counterfactual segregated allocation

This table shows the predicted average test score under both the segregated and random allocation regimes. The segregation dimensions are the main types (except gifted students, as the category is very small). The difference shows the *average counterfactual effect (ACE)*. All effects are demeaned at the level of randomization (school-track for cohorts, school-track-years for classrooms). For each aggregated test score comparison, Panel B shows the variation in the Gini coefficient. For each simulation, 500 random draws are conducted, and standard errors are bootstrapped.



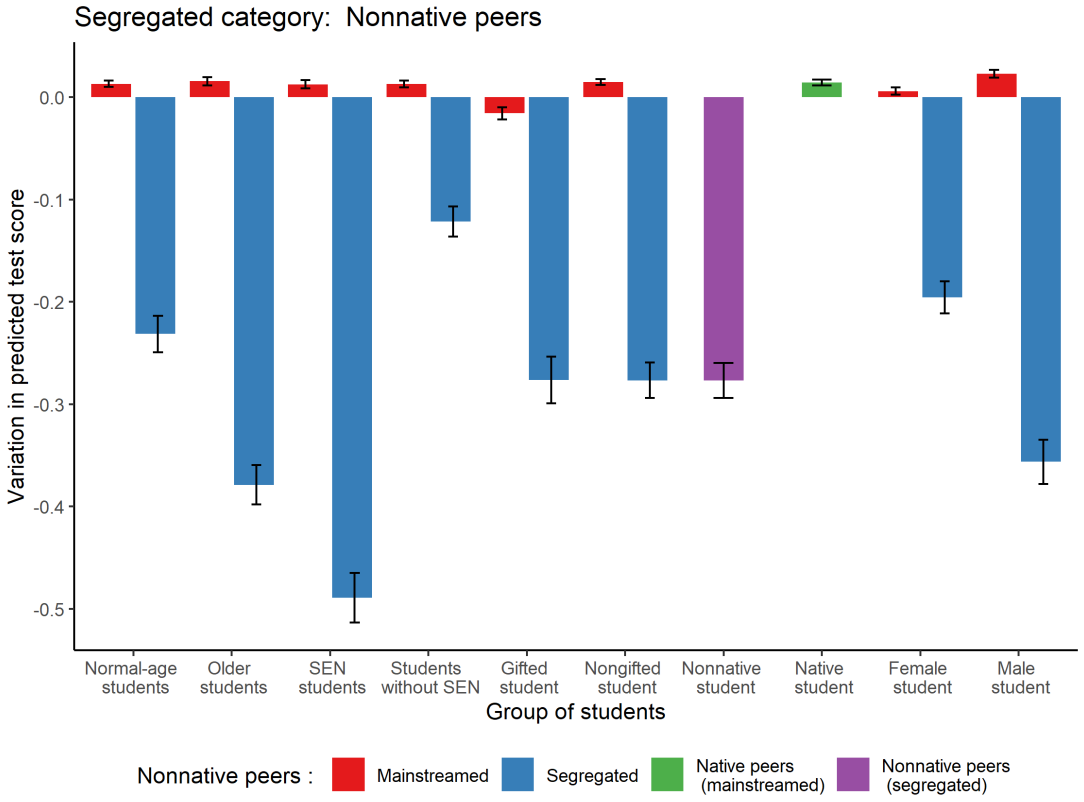

(a)



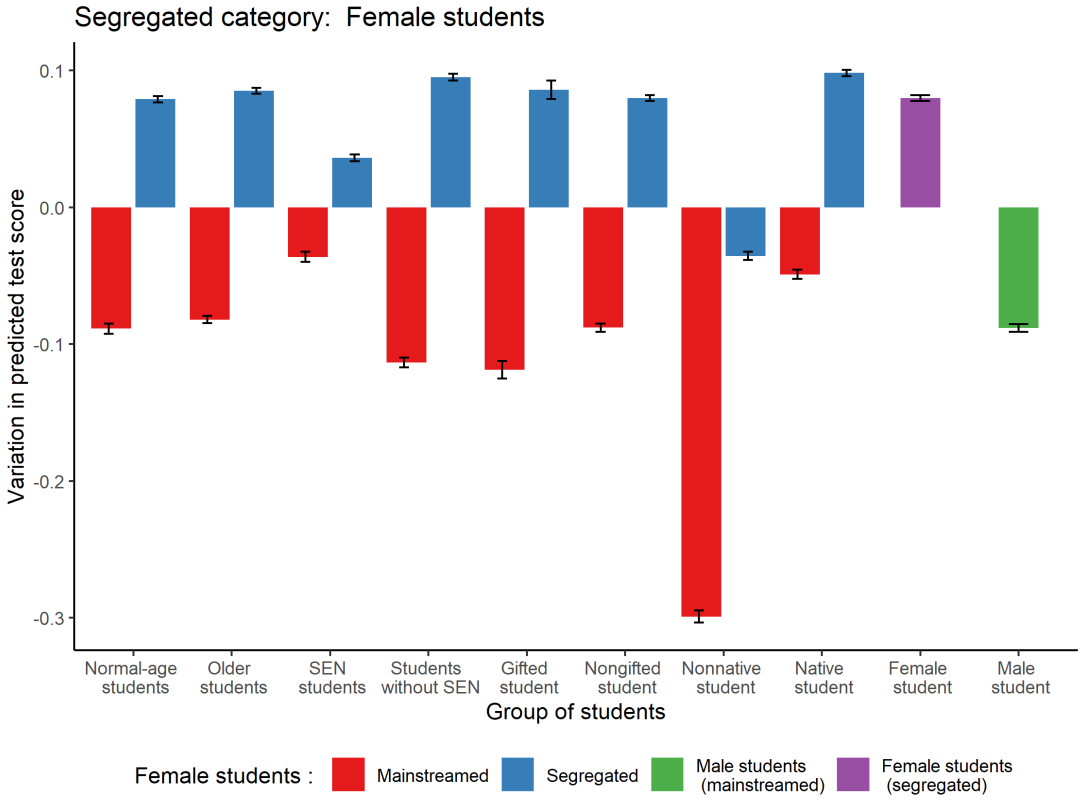

(b)

Figure C.1: CACE: older peers and peers with SEN are segregated

*Notes:* This figure displays the *conditional average counterfactual effects (CACE)* of segregation along the types of older and SEN. "Mainstreaming" means random allocation to classrooms. The green and the purple bars show the own effect, and the red and blue bars show the effects for the other types. Confidence intervals of 95% are obtained by bootstrapping (see main text).




(a)




(b)

Figure C.2: CACE: nonnative and female students are segregated

*Notes:* This figure displays the *conditional average counterfactual effects (CACE)* of segregation along the types of nonnative and gender. "Mainstreaming" means random allocation to classrooms. The green and the purple bars show the own effect, and the red and blue bars show the effects for the other types. Confidence intervals of 95% are obtained by bootstrapping (see main text).

What happens when we segregate students based on their characteristics? For all settings but the setting in which female students are segregated, results follow the same pattern. Everyone in the segregated group is found to be harmed by segregation (blue bars), and this negative impact is always larger than the gains for those who are kept in mixed classes (red bars). Also, the group segregated is usually harmed by segregation (purple bars) in comparison to when they are mainstreamed (green bars). For instance, in Figure C.1a, we see that older male students are the ones suffering the most from segregation by age. In the case of the segregation of older peers, the losses are as high as five times the gains for the mainstreamed group.

Interestingly, segregation along gender generates positive outcomes for segregated female students. However, the gains for female students are almost exactly balanced out by the losses for male students. From a society perspective, mixed education is therefore the best solution, at least when the allocation of resources is kept fixed. These findings corroborate natural experiments exploiting segregation along gender in schools and in tertiary education (e.g., Eisenkopf et al., 2015; Pregaldini, Backes-Gellner, and Eisenkopf, 2020). The only category of students who suffers from gender-segregated environments are nonnative male students. We can only provide speculative interpretation of this: nonnative students in Switzerland mostly come from male-dominated cultures. Gender segregation might exacerbate male-dominated competitive behaviors, in which nonnative students are disadvantaged. What is also striking is the fact that the category of female students who would benefit the most from gender segregated classrooms are gifted female students. This might reflect mechanisms described in the literature about gender differences in competitive behaviors (Niederle and Vesterlund, 2010).

What are the main conclusions of this counterfactual exercise? First of all, all our results strongly suggest that segregation is not a good idea to improve aggregated test scores. This holds even when we incorporate nonlinearities and full heterogeneity in effects. Obviously, segregation is shown to have a positive impact on mainstreamed students, and the strongest improvements in the welfare gains of mainstreamed students happen when we segregate peers with SEN. Second, we show that mainstreaming decreases overall inequality. If we think of education as a public good, and the main mission of public schooling is to give anyone equal chances. Mainstreaming seems to be a good step in this direction. The potential drawbacks of our approach is that, for now, we have ignored group size effects. Moreover, we are aware that segregated classrooms would probably receive additional teacher resources. Thus, our simulation provides lower bounds on the effect of segregation, assuming that resources are constant. But even in this respect, our study is interesting, because we can show, for instance, that a school principal

interested in improving the performance of students with SEN, would have to invest resources such that segregated students with SEN would improve their score by 0.4 standard deviations (average losses of students with SEN in segregation minus the gains of mainstreamed students). In front of such high costs, inclusion seems to be the cheapest option.