# Minorities in Schools -
# Three Empirical Essays in Education Economics

DISSERTATION
of the University of St.Gallen,
School of Management,
Economics, Law, Social Sciences,
International Affairs and Computer Science,
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

**Aurélien Sallin**

from

Villaz-Saint-Pierre (Freiburg)

Approved on the application of

**Prof. Beatrix Eugster, PhD**

and

**Prof. Dr. Ulf Zölitz**

Dissertation no. 5246

Difo-Druck GmbH, Untersiemau 2022

# Minorities in Schools - Three Empirical Essays in Education Economics

DISSERTATION
of the University of St.Gallen,
School of Management,
Economics, Law, Social Sciences,
International Affairs and Computer Science,
to obtain the title of
Doctor of Philosophy in Economics and Finance

submitted by

**Aurélien Sallin**

from

Villaz-Saint-Pierre (Freiburg)

Approved on the application of

**Prof. Beatrix Eugster, PhD**

and

**Prof. Dr. Ulf Zölitz**
**Prof. Dr. Reto Föllmi**

Dissertation no. 5246

Difo-Druck GmbH, Untersiemau 2022

The University of St.Gallen, School of Management, Economics, Law, Social Sciences and International Affairs hereby consents to the printing of the present dissertation, without hereby expressing any opinion on the views herein expressed.

St.Gallen, June 13, 2022

The President:

Prof. Dr. Bernhard Ehrenzeller

*This thesis work is dedicated to my wife, Jen, who has been a constant source of support, inspiration, and encouragement during these wonderful past four and a half years. I am blessed to have you in my life.*

# Acknowledgments

I would first like to thank my supervisor and "Doktormutter", Beatrix Eugster. Beatrix, your insightful guidance, your total trust in my abilities and intuitions, your unconditional support during my PhD adventure, and your drive for purposeful and quality research have helped me grow as an economist, researcher, and as a person. I thank you for your contagious passion for economics, for your constructive and energetic feedback, and for all the opportunities you gave me to further my research.

I would like to thank my colleague and mentor, Simone Balestra. Simone, your pragmatic attitude, creative energy, and passion for academic research have been a great support in navigating the academic world. It's been a pleasure to design the best strategies to gain ground on the minefield of academia, to move through the trenches of desk rejections, and to finally gain victory by making our research accessible to those who might benefit from it. Thank you for this fruitful and fun collaboration!

I would like to thank my colleague and office mate Fanny Puljic. Fanny, I am grateful for your supportive presence over these years, and your commitment to the highest standards have always been an inspiration. Plus, I can't thank you enough for rescuing me many times when I was stranded on the islands of unsolved equations, mysterious algebraic symbols, and dark derivatives.

I would like to thank my senior colleagues for their presence along this PhD path, and for the fertile academic conditions they have provided. A special thank you to Michael Lechner, Stefan Wolter, Helge Liebert, Michael Knaus, Petra Thiemann, Jan Bietenbeck. A big thank you to my dear colleague and fellow cat parent, Caroline Chuard, for this past year at the SEW.

Finally, I would like to thank Mark Schelker. Mark, you not only put in motion my desire to start a PhD in economics, but you actively encouraged me on this path. As the Philosopher would put it, you were instrumental in the actualization of my potentiality. Thank you.

Arbon, June 13, 2022 Aurélien Sallin

# Contents

# List of Figures

# List of Tables

# Executive summary

This thesis empirically investigates three important topics in the literature on Economics of Education: the evaluation of returns to special education programs for students with special needs in inclusive and segregated education settings, the impact of gifted students on their classroom peers, and, finally, the estimation of peer-effects models that flexibly account for the mutual influence of many peer groups.

Chapter 1 addresses the lack of empirical evidence on academic outcomes and labor market returns to special education. I present results from the first ever study to examine short- and long-term returns to special education programs with causal machine learning and computational text analysis methods. I find that special education programs in inclusive settings have positive returns in terms of academic performance as well as labor-market integration. Moreover, I uncover a positive effect of inclusive special education programs in comparison to segregated programs. This effect is heterogeneous: segregation has least negative effects for students with emotional or behavioral problems, and for nonnative students with special needs. Finally, I deliver optimal program placement rules that would maximize aggregated school performance and labor market integration for students with special needs at lower program costs. These placement rules would reallocate most students with special needs from segregation to inclusion.

Chapter 2 investigates the causal impact of intellectually gifted students on their nongifted classmates' school achievement, enrollment in post-compulsory education, and occupational choices. Using student-level administrative and psychological data, we find a positive effect of exposure to gifted students on peers' school achievement in both math and language. This impact is heterogeneous: larger effects are observed among male students and high achievers; female students benefit primarily from female gifted students; effects are driven by gifted students not diagnosed with emotional or behavioral disorders. Exposure to gifted students increases the likelihood of choosing a selective academic track and occupations in STEM fields.

Chapter 3 is an empirical exploration of peer effects in a systematic way. The majority of 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. 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 performance, (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 own academic performance, and that the strongest peer effects are generated by students with special needs, low-achieving students, and male students. Second, we show that classroom peer effects 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 ones that reach maximal aggregated school performance.

# Zusammenfassung

*In dieser Dissertation werden drei wichtige Themen der bildungsökonomischen Literatur empirisch untersucht: die Bewertung der Bildungsrendite von Sonderschulprogrammen für Schüler/-innen mit besonderen Bedürfnissen in integrativen und segregierten Bildungseinrichtungen, die Auswirkungen von begabten Schülern/-innen auf ihre Klassenkameraden/-innen, und schließlich die Schätzung von Peer-Effects-Modellen, die den gegenseitigen Einfluss vieler Peer-Gruppen flexibel berücksichtigen.*

*Kapitel 1 adressiert den Mangel an empirischer Evidenz für die Bildungs- und Arbeitsmarktrendite der Sonderpädagogik. Diese Studie untersucht als erste die kurz- und langfristigen Renditen von sonderpädagogischen Förderprogrammen mit den neusten Methoden des kausalen "Machine Learning" und der "Text Mining" Analyse. Diese Studie zeigt, dass Sonderschulprogramme in integrativen Einrichtungen positive Auswirkungen auf die akademischen Leistungen in Mathematik und Sprache sowie auf Beschäftigung und Löhne von Schülern/-innen mit besonderen Bedürfnissen. Darüber hinaus findet die Studie einen positiven Effekt von inklusiven Sonderschulprogrammen im Vergleich zu segregierten Programmen. Dieser Effekt ist heterogen: Segregation hat den kleinsten negativen Einfluss für Schüler/-innen mit emotionalen oder Verhaltensproblemen und für Schüler/-innen mit besonderen Bedürfnissen und mit nichtdeutscher Muttersprache. Schliesslich liefert diese Analyse optimale Platzierungsregeln, die die aggregierte schulische Leistung und Beschäftigung für Schüler/-innen mit besonderen Bedürfnissen maximieren und die Kosten für die Sonderschulbildung senken. Durch diese Platzierungsregeln würden die meisten Schüler/-innen mit besonderen Bedürfnissen von der Segregation zur Inklusion umverteilt werden.*

*Kapitel 2 untersucht die kausalen Auswirkungen intellektuell begabter Schüler/-innen auf die schulischen Leistungen, die Einschreibung in nachobligatorische Bildungsgänge und die Berufswahl ihrer nicht begabten Klassenkameraden. Basierend auf administrativen und psychologischen Daten auf Schülerebene finden wir einen positiven Effekt des Kontakts mit hochbegabten Schülern/-innen auf die schulischen Leistungen der Klassenkameraden-/innen in Mathematik und Sprache. Diese Wirkung ist heterogen: Größere Effekte werden bei Schülern und Hochbegabten beobachtet; Schülerinnen profitieren in erster Linie von begabten Schülerinnen; die Effekte werden von begabten Schülern/-innen getragen, bei denen keine emotionalen Störungen oder Verhaltensstörungen diagnostiziert wurden. Der Kontakt zu begabten Schülern/-innen erhöht die Wahrscheinlichkeit, dass sich nicht begabte Schüler/-innen für einen selektiven akademischen Weg und Berufe im MINT-Bereich entscheiden.*

*Kapitel 3 untersucht Peer-Effekte auf systematische Weise. Die meisten Studien zu Peer-Effekten im Bildungsbereich haben sich auf die Auswirkungen eines bestimmten Typs von Klassenkameraden/-innen auf Mitschüler/-innen konzentriert. Dieser Fokus ignoriert, dass Peer-Effekte bei Schülern/-innen mit unterschiedlichen Merkmalen heterogen sind und dass es mindestens so viele Peer-Effekt-Funktionen wie Typen von Peers gibt. Diese Studie entwickelt einen empirischen Rahmen, der systematische Interaktionen zwischen Peer-Typen und Nichtlinearitäten von Peer-Effekten berücksichtigt. Methoden des "Machine Learning" werden verwendet, um (i) zu verstehen, welche Dimensionen von Peer-Merkmalen die akademische Leistung am besten vorhersagen, (ii) hochdimensionale Peer-Effekt-Funktionen zu schätzen und (iii) die leistungssteigernde Klassenzimmerzuweisung durch politisch relevante Simulationen zu untersuchen. Erstens stellt die Studie fest, dass die eigenen Merkmale der Schüler/-innen die akademische Leistung am besten vorhersagen und dass die vorhersagbarsten Peer-Effekte von Schülern/-innen mit besonderen Bedürfnissen, leistungsschwachen Schülern/-innen ausgehen. Zweitens zeigen wir, dass die in der Literatur dokumentierten Peer-Effekte wahrscheinlich wichtige Nichtlinearitäten in der Verteilung der Peer-Anteile übersehen. Drittens stellen wir fest, dass eine Klassenzusammensetzung mit möglichst ausgewogenen Schülermerkmalen am besten geeignet ist, um eine maximale aggregierte Schulleistung zu erreichen.*

# 1 | Introduction

> *There is no Algebraist nor Mathematician so expert in his science, as to place entire confidence in any truth immediately upon his discovery of it, or regard it as any thing, but a mere probability. Every time he runs over his proofs, his confidence encreases; but still more by the approbation of his friends; and is rais'd to its utmost perfection by the universal assent and applauses of the learned world.*
>
> — David Hume, *Treatise of Human Nature* (1888)

Since the Salamanca Statement in 1994, most OECD countries have acknowledged that inclusive education is an important means to guarantee equal educational rights for all students (Haug, 2017, p.206).[1] Even though the concept of inclusion is often addressed in relation to the inclusion of students with special educational needs (SEN), an inclusive school system strives to have students of all origins, characteristics, abilities, and from all marginalized groups learn together. In order to successfully include all students in the same learning environment, an inclusive school system has to provide students who would be otherwise segregated in special schools or special classrooms adapted accommodation in the form of special education, and increased individual support from special education professionals. Thus, the implementation of a more inclusive educational system goes hand in hand with the imperative to better understand students' individual needs, how students with particular needs interact with other students in the classroom, and how schools can devise interventions that are the most beneficial to *all* students.[2]

This thesis builds on the conceptual premise that fully understanding inclusive education means understanding the inclusive education production function as a function that accounts for the interplay between all *subgroups* of students, as well as for the

[1] The Salamanca Statement defines inclusive education as follows: "The fundamental principle of the inclusive school is that all children should learn together, whenever possible, regardless of any difficulties or differences they may have. Inclusive schools must recognize and respond to the diverse needs of their students, accommodating both different styles and rates of learning and ensuring quality education..." (p. 11)

[2] The necessity to better understand needs is, for example, well explored in the new research done on students with ADHD (Fletcher, 2014; Fletcher and Wolfe, 2008; Xu et al., 2018) or students with autism

school environment, educational inputs (teachers, resources, remediation and special education programs), and general economic conditions. In particular, this thesis highlights that the notion of inclusion must not be reserved for students with SEN only, but must be extended to students of all characteristics: students on both tails of the intelligence curve (e.g., gifted students), students from other cultural and linguistic backgrounds, students from different socio-economic status, etc. Consequently, the three chapters of this thesis explore inclusive education from three different angles. Starting with the analysis of students with SEN and special education programs in Chapter 2, the focus shifts, in Chapter 3, to the influence of gifted students on their nongifted peers in the classroom. The culmination of this thesis is Chapter 4, in which inclusive schooling is statistically modeled as the product of mutual spillovers across groups of students.

From an economic perspective, the motivation of such research can be summarized as follows. It is crucial to understand school settings and school programs as they are important determinants of skill formation and human capital investment. This is true both with respect to short term skill formation such as academic performance[3], and to long term labor market outcomes.[4] For instance, research in this area is particularly relevant as school settings affect the gender gap in the decision to enroll in STEM careers, an issue frequently discussed in the current literature.[5] Following on this, a thorough understanding of fertile school environments helps policy makers design optimal school environments and best allocate public investment under cost constraints.[6] From this perspective, it is important to understand how students in special education[7] or students with higher ability[8] benefit from, and contribute to, the success of inclusive schooling.

(Autism and Developmental Disabilities Monitoring (ADDM) Network of the CDC).

[3] Carrell, Fullerton, and West (2009); Burke and Sass (2013); Black, Devereaux, and Salvanes (2013); Lavy, Silva, and Weinhardt (2012); Lavy and Schlosser (2011a); Bifulco, Fletcher, and Ross (2011)

[4] Chetty et al. (2011); Anelli and Peri (2017); Fletcher and Wolfe (2008); Fletcher (2014); Mouganie and Wang (2020); Card and Payne (2017); Zölitz and Feld (2021); Anelli and Peri (2017); Black, Devereaux, and Salvanes (2013); Carrell, Sacerdote, and West (2013); Carrell, Hoekstra, and Kuka (2018); Heckman et al. (2010); Heckman, Pinto, and Savelyev (2013); Cunha and Heckman (2007)

[5] Niederle and Vesterlund (2010); Pope and Sydnor (2010); Hyde and Mertz (2009); Nosek et al. (2009); Fryer and Levitt (2010); Card and Payne (2017); Carrell, Page, and West (2010); Brenøe and Zölitz (2020)

[6] Carrell, Sacerdote, and West (2013).

[7] Hanushek, Kain, and Rivkin (2002); Lavy and Schlosser (2005); Keslair, Maurin, and McNally (2012); Schwartz, Hopkins, and Stiefel (2021); Ballis and Heath (forthcoming)

[8] Bui, Craig, and Imberman (2014); Booij, Haan, and Plug (2016); Card and Giuliano (2014).

Chapter 2 addresses the lack of empirical evidence on academic outcomes and labor market returns to special education. In this chapter, I present results from the first ever study to examine short- and long-term returns to special education programs with causal machine learning and computational text analysis methods. I find that special education programs in inclusive settings have positive returns in terms of academic performance as well as labor-market integration. Moreover, I uncover a positive effect of inclusive special education programs in comparison to segregated programs. This effect is heterogeneous: segregation has the least negative effects for students with emotional or behavioral problems, and for nonnative students with special needs. Finally, I deliver optimal program placement rules that would maximize aggregated school performance and labor market integration for students with special needs at lower program costs. These placement rules would reallocate most students with special needs from segregation to inclusion.

Chapter 3 investigates the causal impact of intellectually gifted students on their nongifted classmates' school achievement, enrollment in post-compulsory education, and occupational choices. Using student-level administrative and psychological data, Simone Balestra, Stefan Wolter, and I find a positive effect of exposure to gifted students on peers' school achievement in both math and language. This impact is heterogeneous: larger effects are observed among male students and high achievers; female students benefit primarily from female gifted students; effects are driven by gifted students not diagnosed with emotional or behavioral disorders. Exposure to gifted students increases the likelihood of choosing a selective academic track and occupations in STEM fields.

Chapter 4 is an empirical and systematic exploration of peer effects. The majority of 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. Simone Balestra and I develop a general empirical framework that accounts for systematic interactions between peer types and nonlinearities of peer effects. We use machine-learning methods to understand which dimensions of peer characteristics are the most predictive of academic performance, estimate high-dimensional peer effects functions, and investigate performance-improving classroom allocation through policy-relevant simulations. First, we find that students' own characteristics are the most predictive of own academic performance, and that the strongest peer effects are generated by students with special needs, low-achieving students, and male students. Second, we show that classroom peer effects 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 ones which reach maximal aggregated school performance.

The overall approach of these three chapters is empirical in nature, and attempts to flesh out relationships between *causes* and *effects*. Detecting causal patterns in such a complex environment is difficult, and thus calls for careful identification strategies as well as econometric tools that are effective at expanding the frontiers of our scientific understanding about students in inclusive schooling. Chapter 2 and Chapter 3 are concerned with particular student populations and their effects (students with SEN, and gifted and nongifted students), Chapter 4 takes a more holistic view on inclusive schooling. Finally, these analyses are not only meant to promote our scientific understanding, they are also intended to support policy makers and school administrators in helping *all* students achieve their potential.

# 2 | Estimating returns to special education: Combining machine learning and text analysis to address confounding[*]

Aurélien Sallin (University of St. Gallen)

*Leveraging unique insights into the special education placement process through written individual psychological records, I present results from the first ever study to examine short- and long-term returns to special education programs with causal machine learning and computational text analysis methods. I find that special education programs in inclusive settings have positive returns in terms of academic performance as well as labor-market integration. Moreover, I uncover a positive effect of inclusive special education programs in comparison to segregated programs. This effect is heterogeneous: segregation has the least negative effects for students with emotional or behavioral problems, and for nonnative students with special needs. Finally, I deliver optimal program placement rules that would maximize aggregated school performance and labor market integration for students with special needs at lower program costs. These placement rules would reallocate most students with special needs from segregation to inclusion.*

Keywords: returns to education, special education, inclusion, segregation, causal machine learning, computational text analysis
JEL Classification: H52, I21, I26, J14, C31, Z13

[*] I am grateful to my supervisor Beatrix Eugster, as well as Elliott Ash, Simone Balestra, Brianna Ballis, Uschi Backes-Gellner, Caroline Chuard, Michael Knaus, Edward Lazear, Michael Lechner, Helge Liebert, Bryan S. Graham, Fanny Puljic, Ludger Woessmann, participants of the Law, Economics, and Data Science Group seminar at the ETH, the Causal Data Science Meeting at Maastricht University & Copenhagen Business School, the Machine Learning in Labor, Education, and Health Economics Workshop at the IAB, the EffEE Workshop on Causal Analyses of School Reforms at the WZB in Berlin, the Brown Bag seminar at the University of St. Gallen, the Young Swiss Economists Meeting 2021, the Spring Meeting of Young Economists 2021, the International Conference on Econometrics and Business Analytics (iCEBA) 2021 for their constructive comments and suggestions on early drafts of this paper. I acknowledge financing from the Swiss National Science Foundation (grant no. 176381). This paper reflects the views of the author alone. The usual disclaimer applies.

## 2.1 Introduction

A growing number of students in OECD countries are identified with special needs (SEN)[9]. Taking the US as an example, 14.1 percent of US public school students received Special Education services in 2018–2019, compared to 13.3 percent in 2000-2001, and 10.1 percent in 1980-1981 (NCES, 2020). At the same time, the inclusion of students with SEN in mainstream education has been set as an educational objective by developed countries since the late 1990's.[10] To this end, most OECD countries have reduced segregation of students with SEN and developed a variety of services such as alternative teaching methods and curricula, Individualized Education Programs (IEPS), and increased staff to accommodate individualized support within mainstream education.[11]

Despite the growing number of students with SEN and the increasing implementation of inclusive education programs, empirical evidence regarding how special education (SpEd) placements affect academic and labor market outcomes for students with SEN remains scarce. Existing research shows inconclusive effects of SpEd on academic performance (Hanushek, Kain, and Rivkin, 2002; Lavy and Schlosser, 2005; Keslair, Maurin, and McNally, 2012; Schwartz, Hopkins, and Stiefel, 2021) and on educational attainment (Ballis and Heath, forthcoming). Since (early) interventions in children's school curricula have a profound impact on children's academic and lifelong prospects (see among others Cappelen et al., 2020; Heckman, Pinto, and Savelyev, 2013; Duncan and Magnuson, 2013; Chetty et al., 2011), it is crucial to provide teachers, parents, and policy makers with insights into which SpEd programs are effective. These insights should also help them allocate the most efficient interventions to the students who would benefit the most. This is important in light of the considerable additional financial costs SpEd programs generate for public schools in comparison to standard education (Duncombe and Yinger, 2005; Elder et al., 2021).[12]

[9] Following ICD-10 diagnosis guidelines, students with "special-needs" (SEN) are students suffering from learning impairments, behavioral, emotional or social disorders, communication disorders, physical or developmental disabilities.

[10] According to the United Nations Convention on the Rights of Persons with Disabilities (2006), "States Parties recognize the right of persons with disabilities to education. With a view to realizing this right without discrimination and on the basis of equal opportunity, States Parties shall ensure an inclusive education system at all levels".

[11] See Schwab (2020), and the following OECD reports: "Students with Disabilities, Learning Difficulties and Disadvantages" (OECD Publishing, Paris, 2005), and "TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners" (OECD Publishing, Paris, 2019).

[12] For reference, an annual total of $40 billion was spent exclusively on SpEd in the USA for the 2015 academic year (Elder et al. 2021; NCES, 2015), and educating a student in SpEd can cost twice to three

In this study, I set out to investigate returns to SpEd on academic performance, labor participation, use of disability insurance, and wages. I analyze returns to SpEd in a comprehensive way, and assess returns of six different SpEd programs. The first four programs are offered in inclusive academic settings, and are comprised of counseling, academic support (or tutoring), individual therapies (such as speech therapy), inclusion (students with SEN are mainstreamed but with additional support by a SpEd teacher). Two programs are offered in segregated settings, i.e., semi-segregation (same school but special classrooms), and full segregation (separate schools). In addition, I assess whether inclusive programs are more efficient than segregated programs in generating positive academic and labor market outcomes. I use student-level administrative data on school performance on a compulsory standardized test and social security administrative records. These data are combined with detailed information and psychological written records on each individual student, uniquely linking students' school performance, labor market integration and written psychological assessments for ten consecutive cohorts of students with SEN enrolled in SpEd in the Swiss State of St. Gallen.

I conduct these analyses in the context of the Swiss education system, an academic context which is similar to most OECD countries and the US in terms of inclusive SpEd structures (De Bruin, 2019). Moreover, the Swiss academic setting offers ideal conditions for the investigation of returns to SpEd. A first ideal feature is that the diagnosis of special needs and the SpEd placement decision are conducted by the School Psychological Service (SPS), an external and independent administrative entity. This ensures that treatment is given by professional psychologists, rather than by parents, teachers, or schools. In addition, each school is free to implement the SpEd programs of its choice from a catalogue of measures provided by the Education Ministry. In practice, this means that there is variation in program assignment across schools which is not explained by student or by school characteristics. This is reinforced by the fact that schools in Switzerland were strongly encouraged to implement inclusive programs instead of segregated programs after the Swiss Equality Act for People with Disabilities was passed in 2004. However, not all schools started replacing segregated programs with inclusive programs at the same time. All in all, these features allow me to observe students with similar characteristics and similar SEN but assigned to different SpEd programs.

times as much as educating a mainstreamed student. As an example, the State of California estimates that a SpEd student each year costs $26,000, compared to $9,000 for a mainstreamed student (Overview of Special Education in California report, LAO, 2019).

Special education programs are difficult to evaluate because SpEd placement is based on students' characteristics that are usually unobservable to the econometrician. To tackle the problem of potential selection into SpEd programs, I leverage data that offer unique and unprecendented insights into the placement process: for each student with SEN, I observe all the psychological records and session transcripts written by the caseworkers in charge of both diagnosing the student's special needs and assigning the student to treatment. These records allow me to gain a deep understanding of the students' background, and the nature and complexity of their special needs. To make use of the information contained in text, I implement newly developed techniques in computational text analysis and natural language processing (NLP) adapted to a causal framework (Gentzkow, Kelly, and Taddy, 2019; Mozer et al., 2020; Roberts, Stewart, and Nielsen, 2020; Egami et al., 2018; Keith, Jensen, and O'Connor, 2020). Leveraging text information with methods from the small but steadily growing literature on Double Machine Learning for flexible program evaluation (see, for instance, Chernozhukov et al. (2018); Athey and Wager (2019); Davis and Heller (2017), Knaus, Lechner, and Strittmatter 2020), I am able to account for confounding in unprecedented detail and to plausibly assume unconfoundedness for identification of returns to programs. This study is among the first studies to take advantage of computational text analysis, causal inference, and Double Machine Learning methods to evaluate education programs and returns to education.

I compare students assigned to various SpEd interventions in a pairwise manner (comparing programs that are the most similar, from the most to the least inclusive), as well as students assigned to SpEd interventions with students that were referred to the school psychological service for assessment but not treated. I find that, among all SpEd programs, inclusive programs pay off: first, returns to SpEd programs provided in mainstream education are mostly positive or null in comparison to being referred but receiving no SpEd. I present evidence that targeted individual therapies (such as speech therapy, dyslexia therapies, etc.) are effective at treating preexisting learning disabilities. Moreover, returns to inclusive education in comparison to segregated programs are strongly positive: students with SEN who remain in the mainstream classroom perform better at school, are more likely to participate in the labor market and earn a 15 percentage points higher salary on average than students with SEN segregated into small classes. By conditioning on all the information psychologists report when assigning treatment through written records, I compare students that are similar in all their observed characteristics. On average, I find that estimates based on both covariates and text information are 29% smaller in magnitude than estimates that do not leverage the text information. Moreover, my study suggests that students with SEN who exhibit "disruptive" tendencies (e.g., Lazear, 2001; Carrell, Hoekstra, and Kuka, 2018), i.e., students with social and emotional problems, psychological problems, and nonnative speakers with SEN, are the students who would benefit the most from semi-segregation in comparison to inclusion. Finally, my results highlight that the magnitude of returns vary greatly with the type of program, and thus that sound evaluation of SpEd should account for program specificities.

Getting insights from the literature on statistical treatment rules (e.g., Kitagawa and Tetenov, 2018; Manski, 2004), I further explore optimal policy allocations to inclusive and segregated SpEd programs and make placement recommendations to reach higher aggregate school performance and improve on labor market integration. I propose a set of optimal policies using machine learning algorithms (Athey and Wager, 2021; Zhou, Athey, and Wager, 2018) and compare implemented policies with optimal policies in terms of costs and outcomes. By implementing my proposed optimal policies, a policy maker could significantly increase average school performance and, to a lesser extent, labor market integration at lower overall costs. Easily implementable policies would send all segregated students to inclusion, while more refined policies suggest to keep students with social and emotional problems as well as nonnative students with SEN in semi-segregated settings. I further conduct welfare computations to see whether including students who were previously segregated in the classroom would harm mainstreamed students. I integrate the findings of the quasi-experimental study from Balestra, Eugster, and Liebert (forthcoming) using the same dataset to my analysis, and I find that my optimal policies would generate negligible negative effects on mainstreamed students while significantly increasing average school performance of the reallocated students with SEN.

The present paper contributes to the understanding of returns to SpEd programs. Most studies investigate SpEd as a single, all-encompassing treatment intervention, and compare students in SpEd with students outside of SpEd. Given that SpEd is usually a multifaceted intervention with programs that differ in quality and intensity, these studies fail to provide insights into the effectiveness of different types of programs.[13] Studies have shown moderate effectiveness of SpEd considered as a single program on the academic performance of SEN students (Schwartz, Hopkins, and Stiefel, 2021;

[13] As exceptions, Lavy and Schlosser (2005) and Lovett et al. (2017) focus on targeted remedial education only, and Blachman et al. (2014) look at reading remediation.

Keslair, Maurin, and McNally, 2012; Harrison et al., 2013; Lavy and Schlosser, 2005)[14] and positive returns for SEN students with learning and/or emotional disabilities (Hanushek, Kain, and Rivkin, 2002). In addition, evidence on the effects of SpEd on high-school graduation rates are found to be both positive (Ballis and Heath, forthcoming) and negative.[15] Studies on the effects of SpEd on labor market integration are quasi nonexistent (to the exception of McGee, 2011; Kirjavainen, Pulkkinen, and Jahnukainen, 2016). To my knowledge, this is the first study that assesses short- and long-term returns to SpEd programs at a granular level by ordering SpEd interventions according to their scope and intensity.

Moreover, this study expands on insights from the literature about the factors influencing the emergence of special needs and leading to referrals to SpEd interventions. Many studies highlight the fact that assignment to programs depends heavily on confounders that together influence identification of SEN, assignment to treatment, and the investigated outcomes. For instance, students from non-resilient, low-SES family backgrounds are more likely to develop SEN and to be referred to SpEd (Case, Lubotsky, and Paxson, 2002; Currie and Stabile, 2003; Smith, 2009; Kvande et al., 2018). Other factors that influence referrals include starting school earlier (Balestra, Eugster, and Liebert, 2020; Elder, 2010), racial or ethnic background (Elder et al., 2021) or suspicion of intellectual giftedness (Balestra, Sallin, and Wolter, forthcoming). In this paper, I am able to explore many of these confounders by leveraging individual written psychological records and background information about each student with SEN. Furthermore, I use all this information not only to investigate heterogeneities in returns to programs, but also to devise placement rules that are welfare increasing for all students, and cost-reducing for school officials.

Lastly, this study contributes to investigations of the effects of inclusion in comparison to segregation. On the one hand, existing research offers inconclusive results

[14] Scruggs et al. (2010) conduct a meta-analysis and find overall positive effects of remediation interventions for students with disabilities. SpEd has been shown to have negative or no effects on reading skills, mathematics skills and behavior of SEN students in comparison to non-SEN students in the US (Morgan et al., 2010; Dempsey, Valentine, and Colyvas, 2016). Similar results are documented for Norway (Kvande et al., 2018; Lekhal, 2018), but with positive impact on math skills development. Early preschool SpEd has also been shown to have little to no effects on reading and mathematics skills (Sullivan and Field, 2013; Kohli et al., 2015; Judge and Watson, 2011; Morgan, Farkas, and Wu, 2009).

[15] McGee (2011) for the US and Kirjavainen, Pulkkinen, and Jahnukainen (2016) for Finland report that SEN students have a higher high-school graduation rate than their cognitively equivalent non-SEN peers due to more lenient graduation rules, but lower college enrollment, lower employment rates, and lower wages. Blachman et al. (2014) document that the effects of a randomized reading intervention fade out 10 years after completion of the program.

on the short-term and long-term impacts of inclusion for SEN students (Freeman and Alkin, 2000; Cole, Waldron, and Majd, 2004; Sermier-Dessemontet, Benoit, and Bless, 2011; Daniel and King, 1997; Peetsma et al., 2001; Eckhart et al., 2011)[16] On the other hand, inclusion is reported to have negative effects on peers without SEN in the mainstream classroom (Balestra, Eugster, and Liebert, forthcoming; Rangvid, 2019; Fletcher, 2009). This study bridges the gap between these two strands of literature by investigating in more detail the short-term and long-term impacts of inclusion from the perspective of SEN students, and by investigating optimal inclusive policy rules.

## 2.2 Background and Data

### 2.2.1 Institutional background: special education programs

The implementation of SpEd policies in Switzerland is conducted independently by each Swiss federal state ("canton"). To foster inclusion, the Swiss Equality Act for People with Disabilities (2004) made the equality of access to education for SEN students a priority, and emphasized the promotion of inclusion of SEN students in the main classroom rather than segregation. Thus, inclusion is promoted as the main SpEd intervention tool (Wolter and Kull, 2006), and as a direct substitute for segregation in small special needs classrooms (semi-segregation) (Häfeli and Walther-Müller, 2005). As a result, the share of students sent to segregated schooling has decreased since the Equality Act, while the share of students sent to inclusive schooling has increased. According to the European Agency Statistics on Inclusive Education (EASIE, 2014, 2018), the enrollment rate in mainstream education in Switzerland is similar to other European countries. However, the share of segregated Swiss students with SEN varies substantially across Swiss cantons. The Canton of St. Gallen ranked 5th as the canton with the most segregated SEN students (3.33% of the overall student population vs. 1.85% in Switzerland) in 2010.[17]

[16] In comparison to segregated SEN students, SEN students in inclusive education perform as well in mathematics and even better in literacy (Sermier-Dessemontet, Benoit, and Bless, 2011), exhibit lower motivation but better math performance (Peetsma et al., 2001). However, Daniel and King (1997) find that mainstreamed students with SEN generate more behavioral disruptions, exhibit lower self-esteem, and marginally improve in academic performance. Eckhart et al. (2011) reports that segregated students are less likely to be integrated in the job market and have smaller social networks than students in inclusive environments. The attitude of teachers towards inclusion is also a major influential factor of success for inclusive schooling (Avramidis and Norwich, 2002; De Boer, Pijl, and Minnaert, 2011).

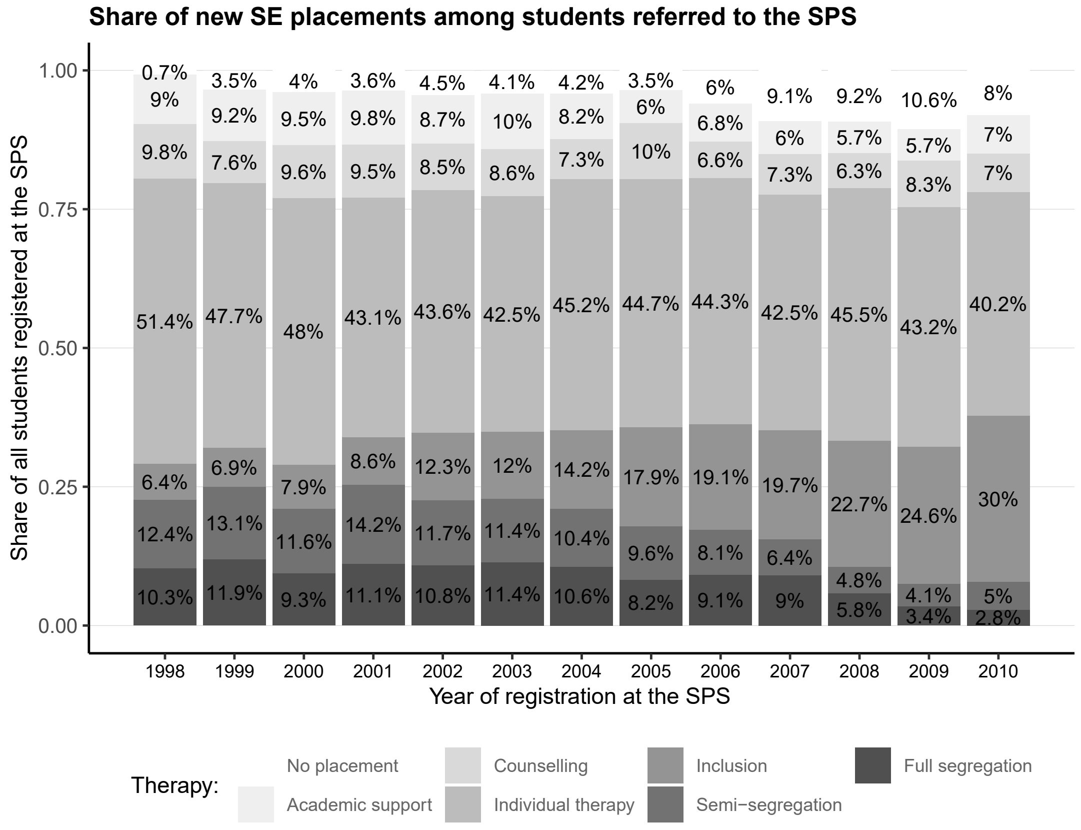
[17] Canton St. Gallen, *Nachtrag zum Volkschulgesetz 2013*, p.38.

This study focuses on students enrolled in SpEd during their mandatory schooling in the Swiss Canton of St. Gallen (around 6% of the Swiss population). The St. Gallen Ministry of Education defines a catalogue of SpEd measures and programs for SEN children.[18] These measures are counseling, academic support, individual therapies, inclusion, semi-segregation, and full segregation. Counseling refers to traditional visits to a therapist or a counselor in which the student's difficulties in school or at home are discussed. It is mostly offered by therapists outside of the School Psychological Service (SPS). Academic support refers to tutoring for children needing additional support for their homework or for learning. Individual therapies are one-to-one or small group sessions; they typically take place during class time, and they target particular learning disabilities for which a particular treatment is required (such as speech therapy, dyslexia or dyscalculia therapy). The inclusion measure refers to all students who received individual inclusive SpEd. These students are provided adapted and goal-oriented complementary teaching by a SpEd teacher who works in the main classroom alongside the main teacher. Semi-segregation refers to small classes (with 10 to 15 students) within the main school. Both inclusive SpEd and semi-segregation are targeted at students with learning and social disabilities, special diagnoses (such as autism, dyslexia, etc.) as well as students who fall behind the class schedule. Full segregation refers to schooling in special schools and targets students for whom mainstream schooling is too challenging (e.g., students with severe disabilities or students suffering from physical impairments such as deafness). Finally, I also observe students who were referred to the SPS for diagnosis but who were not assigned to any treatment ("No placement").

Figure 2.1 displays the newly assigned SpEd interventions in St. Gallen per year. The most frequently assigned therapies are individual therapies. The number of students newly assigned to inclusive SpEd increased from around 6% of students referred to the SPS in 1998 to around 30% in 2010, whereas the number of students assigned to small classes steadily decreased (from 12.4% to around 5%). These figures reflect the actual number of students with SEN being taught in a semi-segregated setting at the primary level ("stocks"), which dropped from 9.17% in 1999 to 6.4% of all students with SEN in 2009, as documented by the official placement register data.

The St. Gallen setting offers many advantages to estimate returns to SpEd programs. First, the diagnosis and SpEd placement decision of students with SEN is

[18] As elaborated in the official document "Kantonales Konzept fördernde Massnahmen" in 2006 by the Canton of St. Gallen, the basic offer includes "SE, speech therapy, rhythm therapy, psychomotor



*Notes:* This figure displays the newly assigned special education interventions per year. It gives the share of students assigned to a particular SpEd program among all students referred to the School Psychological Service over the years. *Source: SPS.*

Figure 2.1: Share of Special Education placements among students referred to the School Psychological Service over the years

conducted by the School Psychological Service, which is an external and independent administrative entity. Therefore, diagnoses and placement decisions are made by SPS psychologists, rather than by parents, teachers, or school administrators. The SPS is organized in eight regional offices. The main task of the SPS is to independently provide diagnoses of learning disabilities, behavioral difficulties, and developmental deficiencies. It assigns therapies and treatments, and offers counseling to students, parents and teachers. As part of the diagnoses, an intelligence test (IQ test) is often administered. After the first consultation, the caseworker, in agreement with parents and teachers, assigns the student to the necessary program. For most students (about nine out of ten), services of the SPS are requested directly by the teacher and/or school official, but some requests are also filed by the parents or the child's medical doctor. Most of

therapy, therapy for dyslexia and dyscalculia, tutoring, special classes". I describe all the therapies given in the canton of St. Gallen in Table A.1.

the requests to the SPS are made when the student is in Kindergarten/Preschool (see Figure A.2).[19]

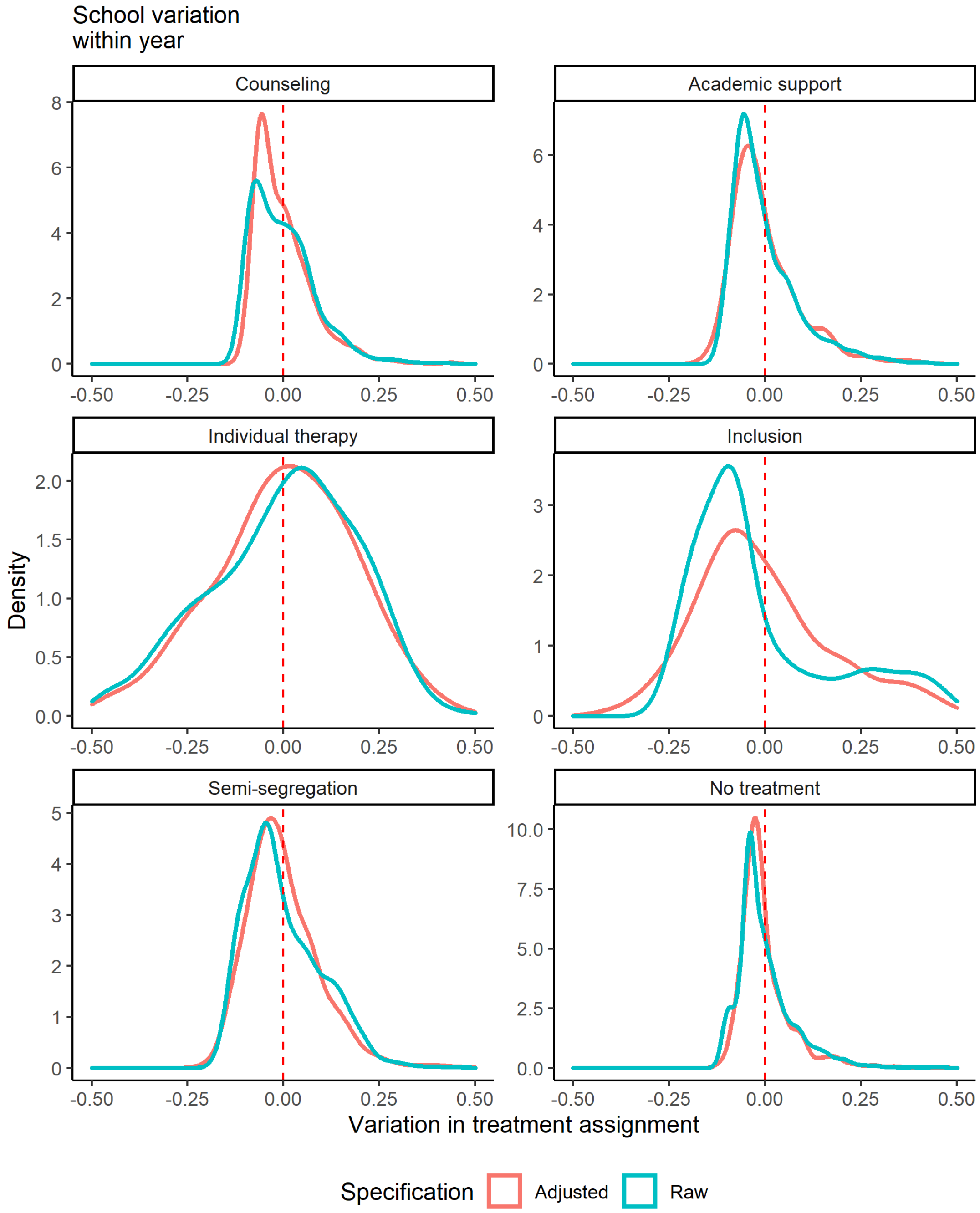
Second, each school is in charge of setting up their own SpEd policies. This offers valuable variation in program assignment within years across schools which is not explained by the students' characteristics. Schools choose, on a yearly basis, which programs to offer among the programs in the catalogue of interventions provided by the Canton. Schools vary substantially in the therapies they offer, as well as in the extent to which they implement inclusive schooling. Figure 2.2 shows the distribution of deviations in the assignment rate of students assigned to each SpEd program per school-year from the mean year inclusion assignment rate for the population of students with SEN. The figure shows that there is substantial variation in assignment to each program across schools within the same year, and that this variation in program assignment is not fully explained by students' and schools' characteristics such as socio-economic score and per-student expenditure (regression-adjusted mean assignment). For instance, some schools have a probability to assign students to inclusion which is more than 50 percentage points higher than the mean assignment to inclusion in the same year. This is valuable information, especially given that students in St. Gallen are assigned to schools on the sole basis of their location of residence. Parents and students must comply with the assignment to the treatment offered by the school.[20]

Third, due to the centralized administration and monitoring of SpEd interventions, programs are similar across schools and use comparable educational technology. Moreover, schools have no real budget constraints when it comes to SpEd programs. This prevents strategic program assignment against additional budget, as documented for some US States (e.g., Cullen, 2003). Schools receive a target amount of therapy-hours from the cantonal central administration, which is calculated on the basis of their "socio-economic score".[21] Within this given amount of therapy-hours, schools

[19] The end of Kindergarten is the moment when teachers decide whether the student is ready for primary school or whether the student needs to take a bridge year. This is in line with Greminger, Tarnutzer, and Venetz (2005), who report that most segregation decisions happen in Kindergarten in Switzerland.

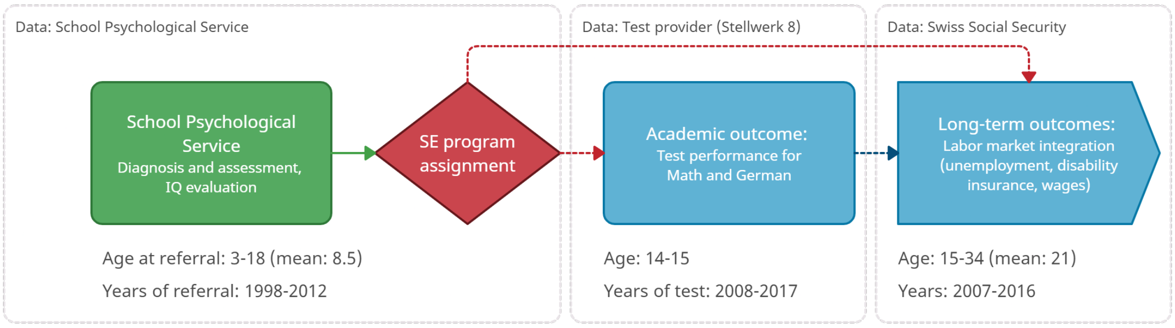
[20] This strict assignment procedure is thoroughly implemented, such that parents have no say about their child's school other than moving permanently to a different municipality or enrolling their students in a private school. Private schooling remains uncommon in Switzerland: in 2014, around 95% of students attend public-funded schools of their community of residence (Wolter and Kull, 2014).

[21] The school socio-economic score is based on the following four indicators: ratio of foreigners with citizenship of non-German-speaking countries in the population group of 5-14-year-olds, share of unemployed in the 15-64-year-old permanent resident population, ratio of 5-14-year-olds dependent on



*Notes:* This figure shows the distribution of deviations (residuals) in the assignment rate of students assigned to all SpEd programs per school-year from the mean year assignment rate for the population of students with SEN. Both the raw deviation and the regression-adjusted deviation are displayed. The adjusted deviation shows deviation adjusted for student-level and school-level covariates. Student-level covariates include gender, nonnative status, IQ, reason for referral, and who sent the student for referral (Panel A of Table 2.1). School-level covariates are shown in Panel B of Table 2.1 and include share of nonnative speakers, share of students with SEN, school size, school socio-economic composition, and urban status. Full segregation is not represented. *Source: Pensenpool, SPS.*

Figure 2.2: Distribution of school-years deviations in assignment to SpEd from mean year inclusion assignment rate


Data: School Psychological Service
Data: Test provider (Stellwerk 8)
Data: Swiss Social Security
School Psychological Service
Diagnosis and assessment, IQ evaluation
SE program assignment
Academic outcome:
Test performance for Math and German
Long-term outcomes:
Labor market integration (unemployment, disability insurance, wages)
Age at referral: 3-18 (mean: 8.5)
Years of referral: 1998-2012
Age: 14-15
Years of test: 2008-2017
Age: 15-34 (mean: 21)
Years: 2007-2016

*Notes:* This figure presents the sample structure as a timeline. All students referred to the School Psychological Service (SPS) between years 1998 to 2012 are observed and receive a treatment. Students' academic performance is observed in the test data ("Stellwerk8") for all students reaching the age of 14 or 15 in years 2008 to 2017. Labor market outcomes are observed in the Swiss Social Security Administration (SSA) data for students reaching the labor market in years 2007 to 2016. Because of attrition and the particular data structure, not all students are observed in both the Stellwerk8 and the SSA data (blue arrow).

Figure 2.3: Visualization of the sample structure

are free to allocate SpEd programs according to their preferred strategy. Schools are obliged to satisfy demand, and often offer more hours than the number of allocated hours. Schools also have a duty to report yearly statistics on the number of SpEd hours offered.

### 2.2.2 Data: main variables and summary statistics

The main data source on students in SpEd are the administrative records from the SPS, academic test scores, data on labor market integration provided by the Swiss Social Security Administration (SSA), and data on schools' statistics about SpEd ("Pensenpool"). Figure 2.3 summarizes the dataset structure and gives an overview of the cohorts represented in the sample. In what follows, I discuss in detail each element of the figure.

**Administrative records from the SPS** The administrative records from the SPS provide information on all students referred to the SPS for a clarification/diagnosis interview between 1998 and 2012. They contain information about the student's characteristics, the therapy assigned, the number of visits to the SPS, and the entirety of

social assistance to the 5-14-year-old population, quota of low-income households with 0-13-year-old children. It is provided by Competence Center for Statistics within the Department of Economic Affairs of the Canton of St. Gallen.

the psychological records written by the caseworker. All summary statistics are reported in Table 2.1, and more detailed statistics per treatment status are given in Table A.3 (columns are ordered from the most inclusive program to the least inclusive program).[22]

Students' characteristics are presented in Panel A of Table 2.1. Forty percent of students in the whole sample are female, and 13% do not have German as their mother tongue. The IQ score is available for 73% of the students, mostly for students in later years as IQ testing at the SPS has become more systematic over the years. At an average of 95, sample IQ scores for SEN students are slightly lower than the population average of 100. Students had on average 10.6 contacts with the SPS, and the number of contacts is strongly positively correlated with the intensity of the program (more contacts are needed for students in segregated programs). Age at first registration is almost 9 on average, which coincides with the start of grading for students attending second grade. The (not mutually exclusive) reasons for referral most commonly mentioned are performance and learning problems (89%), and social or emotional problems (21%). Sixty-six percent of all decisions for referrals are made by the teachers together with the parents of the child. Around 13% of students were enrolled in bridge years between Kindergarten and primary school because of slow development or poor school readiness.

The identification of returns to SpEd in this paper relies mostly on the text contained in the student-level psychological records written by caseworkers. The valuable information contained in the text records makes the assignment process observable. For each visit to the SPS, the caseworker in charge of the student documents the visit, reports the discussion, and gives a recommendation for SpEd placement. Most comments are quite detailed and offer a comprehensive picture of the problems addressed in the discussion, such as family background, psychological issues, the diagnoses of the student, and the particularities of the case.

To be used in estimation, text records must be reduced to some usable representation. In the context of this study, psychological text records are modeled with the intention of learning about the assignment process and adjusting for confounding, while remaining as low dimensional as possible to avoid problems of support and of computational complexity. The text representations should map concepts of the students'

[22] Table A.4 in the Appendix gives the Standardized Mean Difference across all treatment states for all covariates.

| | Mean | Sd | Min | Max | N. obs |
|---|---|---|---|---|---|
| **A: Individual and school characteristics** | | | | | |
| Female | 0.407 | | 0 | 1 | 17,822 |
| Foreign language | 0.126 | | 0 | 1 | 17,822 |
| IQ | 94.92 | 11.9 | 41 | 152 | 13,021 |
| IQ measured | 0.730 | | 0 | 1 | 17,822 |
| Birth year | 1995.35 | 4.3 | 1982 | 2003 | 17,822 |
| Had bridge year (intro class) | 0.134 | | 0 | 1 | 17,822 |
| Age at first interview | 8.563 | 2.3 | 3 | 18 | 17,822 |
| Reasons: social and emotional problems | 0.209 | | 0 | 1 | 17,822 |
| Reasons: performance and learning problems | 0.886 | | 0 | 1 | 17,822 |
| Reasons: problems with teachers or school | 0.027 | | 0 | 1 | 17,822 |
| Reasons: not specified | 0.011 | | 0 | 1 | 17,822 |
| Sent by Caseworker | 0.029 | | 0 | 1 | 17,822 |
| Sent by Others | 0.024 | | 0 | 1 | 17,822 |
| Sent by Parents | 0.052 | | 0 | 1 | 17,822 |
| Sent by Parents and teacher | 0.656 | | 0 | 1 | 17,822 |
| Sent by Teacher | 0.237 | | 0 | 1 | 17,822 |
| Total number of SPS visits | 10.587 | 8.6 | 1 | 152 | 17,822 |
| **B: School characteristics** | | | | | |
| Schools: share of nonnative speakers | 0.223 | 0.1 | 0 | 0.59 | 17,812 |
| Schools: share of SEN students | 0.180 | 0.1 | 0 | 1.00 | 17,812 |
| Schools: total number of students | 170.028 | 151.0 | 0 | 1063 | 17,812 |
| Schools: urban | 0.453 | | 0 | 1 | 17,812 |
| Schools: socio-economic score | 0.980 | 0.1 | 0.79 | 1.20 | 17,812 |
| Schools: per-student expenditure (2017, std.) | 0.00 | 1 | -0.94 | 5.02 | 13,052 |
| **C: Treatment assignment** | | | | | |
| Counseling | 0.081 | | 0 | 1 | 17,822 |
| Academic support | 0.077 | | 0 | 1 | 17,822 |
| Individual therapy | 0.449 | | 0 | 1 | 17,822 |
| Inclusive SE | 0.152 | | 0 | 1 | 17,822 |
| Semi-segregation | 0.095 | | 0 | 1 | 17,822 |
| Full segregation | 0.090 | | 0 | 1 | 17,822 |
| No therapy (but sent to SPS) | 0.056 | | 0 | 1 | 17,822 |
| **D: Outcomes** | | | | | |
| SW8 in SW8 cohort | 0.763 | | 0 | 1 | 13,890 |
| SW8 composit score (SW8 cohort) | 0.000 | 1 | -3.70 | 4.28 | 10,602 |
| Used disability insurance (SSA cohort) | 0.075 | | 0 | 1 | 11,979 |
| Used unemployment insurance (SSA cohort) | 0.234 | | 0 | 1 | 11,979 |
| Monthly wage: last registered year (std., SSA cohort) | 0.000 | 1 | -1.85 | 6.85 | 11,979 |
| **E: Sample attrition** | | | | | |
| In SW8 cohort (1992-2003) | 0.779 | | 0 | 1 | 17,822 |
| In SSA cohort (1982-1998) | 0.672 | | 0 | 1 | 17,822 |
| In both SW8 and SSA cohorts | 0.463 | | 0 | 1 | 17,822 |

*Notes:* Summary statistics for the population of students referred to the SPS in the Canton of St. Gallen. The sample is composed of SN students from the Canton of St. Gallen having visited the SPS between 1998 and 2012. "SW8" refers to students observed in the *Stellwerk8* academic performance dataset, and "SSA" to students observed in the the Swiss Social Security dataset. Standard deviations are not reported for dummy variables. *Source: SPS, SW8, SSA and Pensenpool data.*

Table 2.1: Summary statistics

| **Text Representation** | | **Dimension of covariate matrix** |
|---|---|---|
| **“Bag-of-words”** | *tf* | $N \times 782$ tokens |
| | *tf-idf* | $N \times 921$ tokens |
| | *tf-tf-idf* | $N \times 914$ tokens |
| **Structural Topic Modelling (STM)** | 10 topics | $N \times 10$ topics |
| | 80 topics | $N \times 80$ topics |
| **Topical Inverse Regression Matching (TIRM)** | 10 topics + 1 treatment projection | $N \times 10$ topics |
| **Word2Vec** | 50–dimensional | $N \times 50$ |
| | 100–dimensional | $N \times 100$ |
| **Professional diagnosis** | Dictionary/Keyword approach | $N \times 16$ diagnoses |

*Notes:* This table describes the different Natural Language Processing (NLP) methods for text information retrieval used in this paper. A discussion of these methods, examples and summary statistics can be found in Appendix Section A.1.

Table 2.2: List of used methods for text information retrieval

mental health, learning/behavioral disabilities, and other background information as well as possible; they should also account for the context of words and offer enough nuance to adequately represent the situation of each student. Using text for the purpose of causal analysis to adjust for confounding is a recent enterprise and depends heavily on the empirical setting: there is so far no established standard practice (see relevant discussions in Mozer et al., 2020; Weld et al., 2020; Keith, Jensen, and O’Connor, 2020; Roberts, Stewart, and Nielsen, 2020; Egami et al., 2018).

Table 2.2 summarizes the computational apparatus used to extract information from text. To avoid making estimates too dependent on the choice of text information retrieval method, I extract information from the text using five different state-of-the-art NLP methods and nine different specifications: the term-document matrix (TDM) representation, or “bag-of-words” (see, for instance, Mozer et al., 2020); structural topic modeling and topical inverse regression matching, which learn topics and context of words in a semi-supervised manner (Roberts, Stewart, and Airoldi, 2016; Roberts, Stewart, and Nielsen, 2020; Blei, Ng, and Jordan, 2003); neural network embeddings such as Word2Vec in which words are embedded in a lower-dimensional space (Mikolov et al., 2013); dictionary representations that map professional diagnoses. For each method, the final dimension of the text representation matrix is presented. The features contained in the representation matrix are subsequently used as controls for estimation of treatment effects. I discuss how I implement each of these methods and provide descriptive statistics for each method in Appendix Section A.1.

**Program assignment** SpEd programs of interest are defined as the programs figuring in the cantonal catalogue of measures mentioned in Table A.1. Around 37% of the students were given individual, one-to-one therapy only, such as speech therapy, dyslexia therapy, or dyscalculia therapy. Thirteen percent of all students are placed in inclusive settings, around 16% in segregated settings (8% in semi segregation and 8% in full segregation). Some students (around 5%) were referred by their teachers to the SPS but did not receive any SpEd intervention. These students form an interesting comparison group, since they are students who raised strong suspicion for SpEd referral but who do not receive SpEd placement after all. From the notes, I know that most of these students have been received and assessed by a caseworker who in turn decided that no further intervention was needed.

Even though SpEd interventions are defined at the central level, program effectiveness might vary with schools' characteristics. To account for this, I bring in statistics about schools obtained from the *Pensenpool* data of the Ministry of Education. I use four measures about the school population (share of students with SEN, share of foreign students, total school population, urban or rural school). Moreover, I use two measures of educational inputs previously used in the literature: standardized per-student spending (e.g., Jackson, Johnson, and Persico, 2016)[23], and the "socio-economic score", introduced above, which measures the school's socio-economic composition (e.g., Angrist and Lang, 2004). These school-level statistics are measured in the year in which students with SEN are assigned to treatment (with the exception of per-student spending, which is only measured in 2017). Table A.3 shows that school characteristics are rather well balanced across all treatments, with the exception of inclusion and semi-segregation. Schools implementing semi-segregation tend to be more urban, larger, and with a higher share of foreign students. However, schools offering inclusion and schools offering semi-segregation overlap in their characteristics, as can be seen in Figure A.1.

**Outcomes: test scores and labor-market integration** I measure different outcomes to capture school achievement as well as labor market integration. Outcomes are reported in Panel D of Table 2.1. For academic performance, I use test scores from the

[23] The data on spending per primary school students comes from the official accounts published by municipalities at the end of the fiscal year. According to the data, municipalities spend on average 10,160 Swiss Francs (approximately 11,140 USD) per primary school student. This figure is higher than the OECD average (8,733 USD) but comparable to the corresponding figure in the U.S. (11,319 USD), as the OECD documents (OECD, 2017).

"Stellwerk8" standardized test (SW8) taken in grade 8, which give the individual academic achievement for the entire population of students enrolled in 8th grade during the years 2008 to 2017. This test is mandatory for all students with SEN (except for students in fully segregated settings) and is the same in all schools. It is computer-based, and automatically adapts the difficulty of questions to the ability and knowledge revealed by the student in the previous questions. It tests core knowledge of mathematics, language (German), and, depending on the track, other subjects. I focus on the composite score in German and Math, which are compulsory subjects for all students. Test scores range between 0 and 1,000 (1,000 being the best), and are standardized by school-year for easier interpretation and comparison. The performance on the test is important both for students, who will use the test scores when choosing their post-compulsory education, and for teachers, whose relative performance can be reflected in the rate of success of their students. As students with SEN in fully segregated settings are not required to take the test and can choose to opt out, I create a test-taking indicator variable to account for attrition.

Data on labor market integration are provided by the Swiss Social Security Administration (SSA) for the years 2007 to 2016, and contain the individual history of wages, whether the individual has benefited from a disability insurance status (DI), and whether the individual has requested unemployment insurance. I compute the income as the last income recorded standardized over birth years. This gives the average relative position of individual income per cohort and per year, which accounts for cohort as well as year effects.[24] Income is defined as income from one's own labor, namely net of DI and unemployment benefits. Around 8% of the sample have claimed disability insurance, and 23% have claimed unemployment insurance.

**Sample restrictions** Some restrictions are imposed on the data (details can be found in Table A.2). I discard students who received therapies or measures that are not offered by the schools (for instance, private tutoring). Moreover, I conservatively discard students who received so-called secondary "supportive measures" only[25], and students who received more than one treatment. This ensures that multiple influences of different treatments are not confounding the main treatment.

[24] I also checked other income definitions, such as last monthly income recorded standardized over birth years. Results are robust across these alternative income definitions.

[25] These measures include tutoring, language classes for students with an immigration background, and gifted education. For details, see the "Sonderpädagogik-Konzept" of the Canton of St. Gallen, available on the website of the St. Gallen schools. Students receiving supportive measures in addition

Cohorts registered in the school data and cohorts registered in the SSA data do not perfectly overlap (see the red arrows in Figure 2.3). Since the SW8 test was given in years 2008 to 2017, and given that some cohorts were not exposed to the test, I investigate subsamples for each outcome separately. Subsample sizes are reported in Panel E of Table 2.1. While 78% of the sample were in cohorts subject to the SW8 test, 67% are from cohorts with no test but with recorded labor market outcomes. Finally, 46% of observations are observed in both subsamples. Attrition is only due to cohort variation, and I conduct attrition analyses in my robustness checks to show that attrition is not a problem for my main results.

## 2.3 Empirical strategy

A plausible causal estimation of returns to SpEd programs requires comparing the academic and labor outcomes of students who are similar in all the characteristics which jointly influence their outcomes and their assignment to SpEd programs. In the absence of a randomized experiment in which students are randomly assigned to programs, I leverage the information contained in the psychological reports, and I model the assignment process with a unusually exclusive and detailed perspective. Furthermore, I make implicit use of the exogenous variation in treatment assignment within years across schools to identify effects of SpEd programs.

### 2.3.1 Definition

I compare the outcomes of students assigned to various SpEd interventions in a pairwise manner (from most to least inclusive interventions), as well as students assigned to SpEd interventions with students that were referred to the SPS but who were not treated. I follow a multivalued treatment framework in observational studies (Imbens, 2000; Lechner, 2001), in which I compare program $d$ with program $d'$ for student $i$. More precisely, I denote by $d$ the received treatment by student $i$ among the set of mutually exclusive seven programs $\mathcal{D}$. The observed outcome given $i$'s assigned therapy is $Y_i = \sum_{d=1}^{D} \underline{1}(D_i = d)Y_i^d$, and the potential outcome for each individual is $Y_i^d$ for all $d \in \mathcal{D}$. I further denote as $\mathcal{X}$ a set of pre-treatment variables, and as $\mathcal{Z}$ the subset of $\mathcal{X}$ that contains the variables used to conduct heterogeneity analysis. The general-

to the main measures are, however, kept in the dataset.

ized propensity score is defined as $p_d(x) = P(D_i = d|X_i = x)$, namely the conditional probability of receiving each treatment.

I am interested in the following estimands. The first is the average potential outcome (APO) under each treatment $d$, $\text{APO}_d = E[Y_i^d]$. It is the average outcome for the whole population as if it was assigned to program $d$. This corresponds to the "value" of each program. The second is the pairwise Average Treatment Effect $\text{ATE}_{d,d'} = E[Y_i^d - Y_i^{d'}]$, which represents the effect of treatment $d$ vs treatment $d'$ as if everyone in the population was observed under both treatment states. Since some treatments might not be available for the whole population (for instance, full segregation is not a feasible intervention for all SEN students), the ATE is not interesting for all treatment pairs. I thus compare treatment effects for the subpopulation actually observed in a given program using the Average Treatment Effect on the Treated $\text{ATET}_{d,d'} = E[Y_i^d - Y_i^{d'}|D = d]$. Comparing the ATE and the ATET gives valuable insights about the program assignment process: a large difference between the two estimates might underline effect heterogeneity or nonrandom assignment into programs. Finally, I look at Conditional Average Treatment Effects (CATEs). I consider two different cases of CATEs: first, Group Average Treatment Effects (GATEs) give the ATEs for predefined and policy relevant groups of students, i.e. $\text{GATE}_{d,d'}(z) = E[Y_i^d - Y_i^{d'}|Z_i = z]$ where $Z_i \in \mathcal{Z}$. For instance, I investigate whether treatment effects are heterogeneous for students with and without behavioral problems. Second, I look at Individual Average Treatment Effects (IATEs) for ATEs at the most granular, individual level. Instead of focusing on groups, IATEs include all observed confounders as heterogeneity variables. This is expressed as $\text{IATE}_{d,d'}(z) = E[Y_i^d - Y_i^{d'}|X_i = x]$, where $X_i \in \mathcal{X}$ (i.e. a vector of observed pre-treatment variables).

### 2.3.2 Identification

The previous section presented the estimands of interest as potential outcomes. As each student is only observed in one program, only one potential outcome per student is observable and the other potential outcomes are latent. Therefore, the estimands of interest are not identified unless the following standard assumptions hold (Imbens and Rubin, 2015). The first key identifying assumption is unconfoundedness, i.e. that the vector of observed pre-treatment covariates $X_i$ contains all the features that jointly influence treatment and potential outcomes $Y_i^d \perp\!\!\!\perp D_i|X_i = x, \forall x \in \chi, \forall d \in \mathcal{D}$. The plausibility of this assumption is justified by the use of text information: the information extracted from the text delivers a unique and detailed overview of both pre-treatment information relevant for treatment assignment and details on the treatment assignment itself. To support this assumption, I show that text brings additional information which is richer than the information contained only in covariates not extracted from text. Section A.1 in the Appendix, and more precisely Figure A.1, Figure A.2, Figure A.3, and Figure A.4, provide evidence that the text delivers valuable additional information which is not contained in the covariates not extracted from text. These figures also show how text information is related to treatment assignment. In addition to the use of text, unconfoundedness is particularly plausible in my setting given the unexplained variation in treatment assignment within years across schools.

The second identifying assumption states that confounders are exogenous, i.e. confounding variables in $\mathcal{X}$ (and in $\mathcal{Z}$) are not influenced by the treatment in a way which is related to the outcomes. This assumption would be violated if covariates are measured after treatment assignment. Non-text covariates are measured before treatment assignment, but text covariates require more scrutiny. To avoid text-induced post-treatment bias, I only use records written before the treatment assignment, thereby removing therapy evaluations and reports about the progress of the student. I also strip out of the text all mentions or discussions of interventions *per se*. The exogeneity assumption would also be violated if referrals to the SPS and treatment assignment would be done based on expected treatment returns, i.e. teachers would refer only the students who are expected to benefit from SpEd to the SPS, and psychologists would be perfectly able to predict the outcomes from treatment assignment. On the one hand, this problem is likely mitigated given that SPS centers as well as the organization of SpEd in St. Gallen are centralized, and that there is no budgetary constraints for SpEd. On the other hand, psychologists are not able to perfectly predict outcomes given treatment, since not all treatments are available in all schools and in all years.

Third, overlap (or common support) $0 < p_d(x) < 1, \forall x \in \chi, \forall d \in \mathcal{D}$ ensures that SEN students can be compared at all values of $X$ for a given treatment effect. Because of the variation offered by the school choices of supplied interventions, I can observe students with similar characteristics who were offered different interventions. To make this assumption even more plausible, I compare only interventions that are closest in terms of intensity and in terms of the special needs they target. Moreover, to deal with potential problems of overlap, I present effects for the overlap population in the section A.2. Finally, lack of overlap is problematic for students assigned to fully segregated SpEd programs, since this particular population of SEN students exhibits more severe mental and learning disabilities than other SEN students. This will be kept in mind in the discussion of the results.

Finally, assignment to a particular SpEd program does not generate spillover effects (SUTVA), i.e. $Y_i = Y_i(D_i)$. There are mainly two cases in which SUTVA could be violated. First, the presence of a student with SEN in a program might generate spillover effects. I estimate total effects of programs on the population of SEN students only (and not on the mainstream population). In other words, my estimates incorporate potential classroom spillovers.[26] Second, if therapies are budgeted at the school level, sending one student to therapy might reduce available resources for other SEN students who also need therapy. This is not a concern in this setting. As mentioned above, schools in St. Gallen do not engage in strategic therapy assignment against additional budget (e.g., Cullen, 2003), as there are no real budget constraints when it comes to SpEd.

Under these assumptions, the estimands of interest are identified using the "augmented" weighted estimator (AIPW) score $\Gamma^h_{d,X_i}$. This score combines the conditional expectations of the outcome $Y$ specific to each potential treatment, $\mu(d,x) = E[Y_i|D_i = d, X_i = x]$ with the outcome residual reweighed by some function of the treatment probability $p_d(x)$. Following the "balancing weights" notation of Li and Li (2019), the general form of this estimator is:

$$\Gamma^h(d, X_i) = \mu(d,x)h(x) + \underline{1}(D_i = d)(Y_i - \mu(d,x))\omega_d(x), \tag{2.1}$$

The "tilting function" $h(x)$ defines the target population as a function of the propensity score $p_d(x)$, and $\omega_d(x) = h(x)/p_d(x)$.[27] When the population of interest is the whole population, as in the ATE, the tilting function $h(x)$ is 1 and the estimator is doubly robust (Robins, Rotnitzky, and Zhao, 1994, 1995).[28]

[26] Note that there is no strategic assignment of SEN students to mainstream classrooms in St. Gallen (see Balestra, Sallin, and Wolter, forthcoming; Balestra, Eugster, and Liebert, forthcoming).

[27] Note that $\omega$ can accommodate weights for different subpopulations as additional "balancing weights" schemes, such as trimming weights or matching weights (Li, Morgan, and Zaslavsky, 2018; Li and Li, 2019).

[28] The score is doubly robust when it is still consistent if the propensity score or the outcome equation are misspecified. For more details on the APO score and the double robustness properties, see Glynn and Quinn (2010) for an intuitive introduction and Knaus (2021) for DML.

All estimands of interest mentioned above are identified as follows:

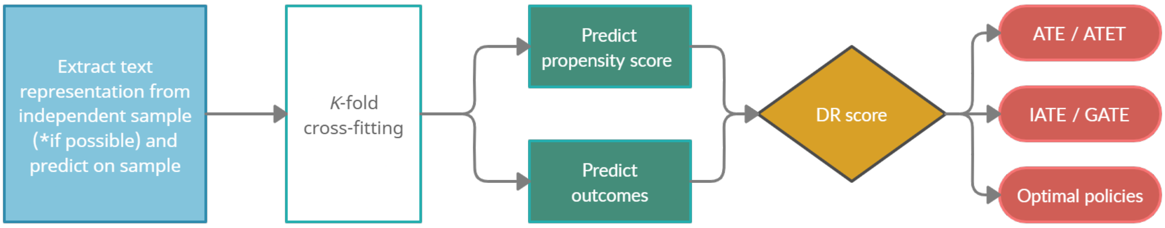
$$\begin{aligned}
\text{APO}_d &= E[\Gamma^h(d, X_i)], & h(x) &= 1 \quad (2.2)\\
\text{ATE}_{d,d'} &= E[\Gamma^h(d, X_i) - \Gamma^h(d', X_i)], & h(x) &= 1 \quad (2.3)\\
\text{ATET}_{d,d'} &= E[\Gamma^h(d, X_i) - \Gamma^h(d', X_i)|D_i = d], & h(x) &= p_{d'}(x) \quad (2.4)\\
\text{GATE}_{d,d',z} &= E[\Gamma^h(d, X_i) - \Gamma^h(d', X_i)|Z_i = z], & h(x) &= 1 \quad (2.5)
\end{aligned}$$

The APO $\Gamma^h(d, X_i)$ for the ATE score takes $h(x) = 1$ since it applies to the whole population. The estimand for the ATET takes $h(x) = p_{d'}(x)$ as it applies to the population of the treated. Note that, in my main specifications, I do not explicitly model the variation in assignment rate across schools within years. These variations contribute to the plausibility of the four assumptions presented above. However, I estimate an IV specification with the within year across school variation in program assignment rate as an instrument in Appendix Section A.2.4.

### 2.3.3 Estimation with Double Machine Learning

The estimation procedure is represented in the stylized workflow of Figure 2.4. In a first step, text representations are extracted from an independent, held-out sample in order to avoid risks of overfitting, and are subsequently predicted on the main sample. This ensures that text representations are meaningful across the whole dataset. Once text representations are predicted on the main sample, $K$-fold cross-fitting (see Chernozhukov et al., 2018) is used to estimate the two nuisance parameters of interest $\hat{\mu}(d, x)$ (estimated conditional expectation of the outcome) and $\hat{p}_d(x)$ (estimated propensity score) using the covariates $\mathcal{X}$ presented in Panel A of Table 2.1 as well as text covariates. Succinctly, the procedure works as follows: (i) the sample is randomly split in $K$ folds of equal size, (ii) one fold is left out, and the remaining $K - 1$ folds are used to train machine learning models to estimate the nuisance parameters. These models (iii) are used to predict $\hat{p}_d(x)$ and $\hat{\mu}(d, x)$ on the left-out $K$th fold, and (iv) the procedure is repeated such that each fold is left out once. Extracting text representations from an independent sample requires the availability of large amount of data, which is not always available. When not available, text representations can be retrieved alongside the training of nuisance functions within each $K - 1$ folds.[29] In this

[29] For each fold, the model used for text representation is trained on the $K - 1$ training folds and is in turn used as a set of covariates to train the predictive model for the propensity score or the outcome. In this case, text representations are discovered in each $K - 1$ fold and thus are fold-dependent. They



*Notes:* This figure represents a stylized workflow of the estimation procedure. First information from text is retrieved, then used in $k-$fold cross-fitting to estimate the two nuisance parameters (estimated propensity score and estimated conditional expectation of the outcome). The doubly-robust score is computed and used to estimate the estimands of interest (APO, ATE, ATET, IATE and GATE, optimal policies).

Figure 2.4: Workflow of Double Machine Learning and text analysis

application, text representations are extracted from the City of St. Gallen sample in the case of Word2Vec and the dictionary. Topics for STM and TIRM are extracted within each $K-1$ folds.

The nuisance parameters are then combined to build, on each left-out fold, the doubly-robust (DR) score as:

$$\hat{\Gamma}^{h=1}_{i,d} = \hat{\mu}(d, X_i) + \frac{1(D_i = d)(Y_i - \hat{\mu}(d, X_i))}{\hat{p}_d(X_i)}. \tag{2.6}$$

Since no observation is used to estimate its own nuisance parameter, cross-fitting reduces the risk of overfitting. I estimate the nuisance parameters with a combination of many methods through an ensemble learner (Van der Laan, Polley, and Hubbard, 2007): I predict the nuisance parameters with three ML methods (Lasso, Elastic Net and Random Forest) and with 11 different text representations on top of main covariates. This results in 33 different estimations per fold. I obtain the weights of the ensemble learner by cross-validating the out-of-sample MSE of each specification and use a weighted combination of the 5 most predictive specifications in the final score.

From the score of the APO defined in Equation (2.6), the ATE is constructed as the mean of the difference between the APO scores for the treatments of interest, i.e. $\widehat{\text{ATE}}_{i,d,d'} = \hat{\Gamma}^{h=1}_{i,d} - \hat{\Gamma}^{h=1}_{i,d'}$. For the ATET, the doubly-robust score for $\widehat{\text{ATET}}_{i,d,d'}$ is $\left[\frac{1(D_i=d)(Y_i-\hat{\mu}_{d'}(X_i))}{\hat{p}_d} - \frac{\hat{p}_d(X_i)}{\hat{p}_{d'}(X_i)}\frac{1(D_i=d')(Y_i-\hat{\mu}_{d'}(X_i))}{\hat{p}_d}\right]$ where $\hat{p}_d = P[D = d] = N_d/N$ (Farrell,

cannot be compared to text representations in other folds, and cannot be used as covariates of interest in the set $\mathcal{Z}$. To reduce computing times, the vocabulary (the set of tokens) is extracted for the whole dataset before cross-fitting.

2015). For point estimates of the APO, ATE and ATET, I take the means of the different estimands and rely on single-sample $t-$tests for statistical inference.[30] The GATEs are estimated by taking the conditional mean of the $\widehat{\text{ATE}}_{i,d,d'}$ over groups determined by pretreatment variables $Z_i$, i.e. by regressing the score $\widehat{\text{ATE}}_{i,d,d'}$ on the group variables of interest and using standard heteroscedasticity robust standard errors (following Semenova and Chernozhukov, 2021). To assess the effect heterogeneity along a continuous variable $Z_i$, Zimmert and Lechner (2019) and Fan et al. (2020) propose to regress the individual score of $\widehat{\text{ATE}}_{i,d,d'}$ on $Z_i$ with a kernel regression and standard inference for nonparametric regression. I estimate second-order Gaussian kernel functions and choose the 0.9 cross-validated bandwidth, as recommended by Zimmert and Lechner (2019). Finally, I estimate IATEs by using a DR-learner, i.e. I train an ensemble learner to predict the individual ATE score $\widehat{\text{ATE}}_{i,d,d'}$ out-of-sample (see Kennedy, 2020; Knaus, 2021).[31]

Alongside its double-robustness property, the use of Double Machine Learning (DML) and of the AIPW score has many advantages when working with text. First, it allows for leveraging text representations both in the propensity score (as in Mozer et al., 2020; Roberts, Stewart, and Nielsen, 2020) and in the outcome equation, which reduces problems of extreme propensity score accuracy (Weld et al., 2020) and overcomes difficulties of matching on both covariates and text.[32] Second, by not relying on one particular estimation method but combining many of them in an ensemble learner, I make full use of different ML methods and use the ones that work best with each text representation. This also mitigates potential misspecification of the text and covariate functional forms.[33]

[30] This is possible without taking into account the fact that nuisance parameters are estimated in the first place if the nuisance parameters estimators are consistent at a relatively fast rate, asymptotically normal and semiparametrically efficient (Chernozhukov et al., 2018).

[31] I follow the following procedure: in a first step, I predict in each fold the nuisance parameters and then compute the individual score $\widehat{\text{ATE}}_{i,d,d'}$. In a second step, I train an ensemble learner in the same folds to predict $\widehat{\text{ATE}}_{i,d,d'}$ from covariates $X$. In a third step, I use the trained ensemble learner to predict $\widehat{\text{ATE}}_{i,d,d'}$ on the left-out fold. This procedure is computationally heavier than an in-sample IATE prediction but has the advantage of avoiding overfitting. It is however less computationally burdensome than the cross-fitting procedure proposed by Knaus (2021), as I do not have to re-estimate, in each fold, the text measures that need to be estimated in-sample.

[32] Matching algorithms for text as proposed by Mozer et al. (2020) are both computationally burdensome and difficult to implement, insofar as assessing match quality of text is difficult (researchers must find the relevant text reduction, the relevant text distance metrics, and the relevant matching assessment tool, such as human coders).

[33] For instance, the *Generalized Random Forest* of Athey, Tibshirani, and Wager (2019) or the *Modified Causal Forests* of Lechner (2019) rely exclusively on random forest, which might not perform well on a "bag-of-words" representation of text due to the high number of sparse dummy variables.

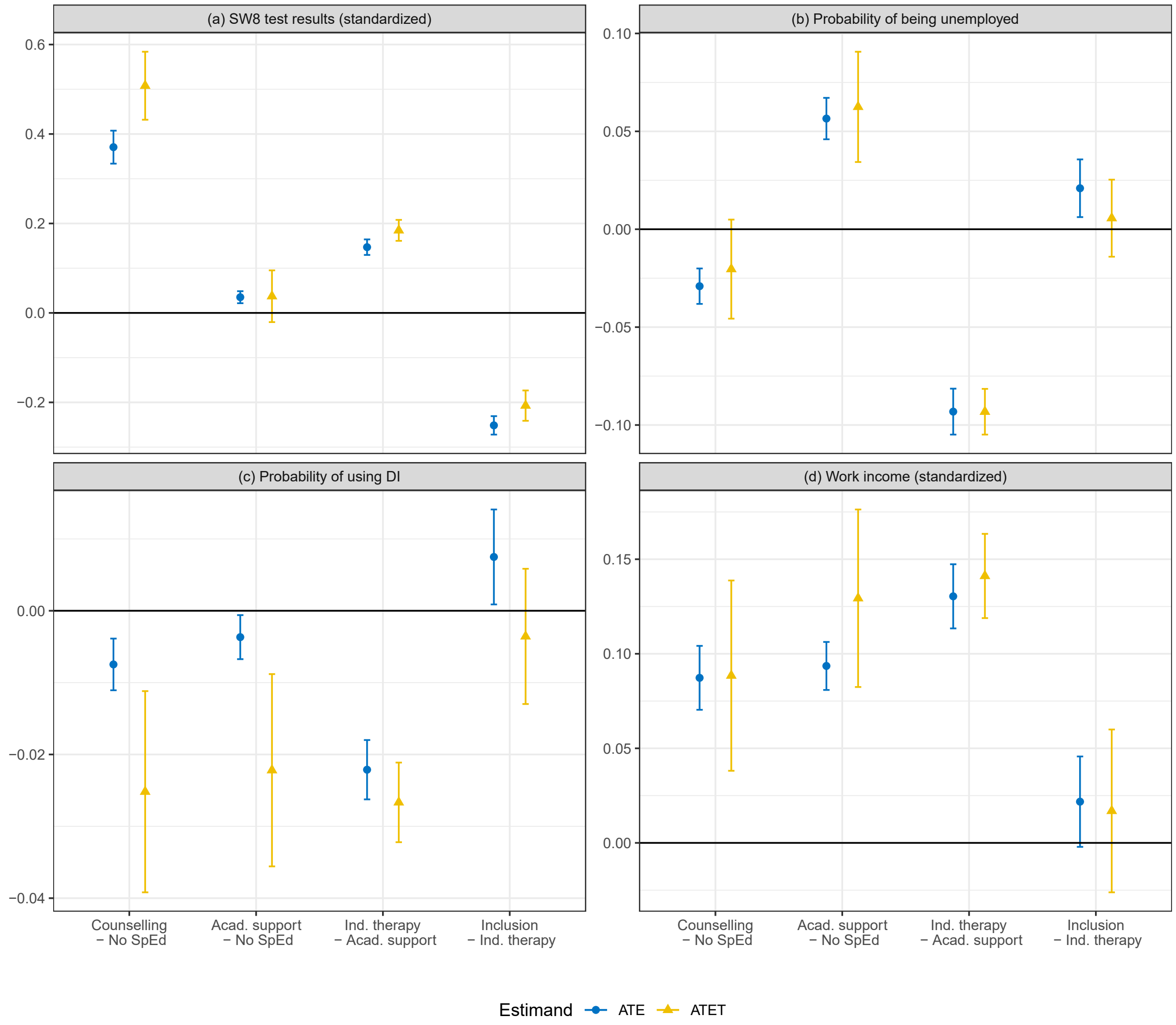
## 2.4 Results: returns to special education programs

In this section, I present different sets of main results: first, I present the pairwise effects for inclusive SpEd interventions (i.e., interventions that are provided in the mainstream school environment). Second, I focus more specifically on the effect of inclusion vs. semi-segregation. Third, in order to relate to existing literature, I look at the "extensive margin" of SpEd interventions and assess the effect of being assigned to a program vs. being assigned to no program at all. This set of results corresponds to the effect traditionally estimated in the literature. Fourth, I conduct analyses of the heterogeneous effect of inclusion. Finally, I perform a series of further analyses and robustness checks.

### 2.4.1 Returns to Special Education programs in inclusive school settings

I first present, in Figure 2.5, returns to SpEd on academic performance for interventions that are the closest in degree of severity and inclusion, and which are either provided as supportive or remediation measures (counseling, academic support or tutoring, individual therapies, and inclusion) provided in the mainstream school environment. Results read as follows: pairwise effects give the effect for being assigned to the first program (for instance, in the first column, to counseling) instead of being assigned to the second program (e.g., to no program) on academic performance (Panel a), probability to be unemployed (Panel b), the probability to use disability insurance (Panel c), and on work income (Panel d). Test scores and wages are standardized with mean 0 and standard deviation 1. The baseline ("No SpEd") probability of unemployment benefit recipiency is 0.19, and 0.07 for disability insurance recipiency. Effects account for all the observed confounding from covariates (such as gender and IQ) as well as all information contained in the psychologists' records. Point estimates and 95% confidence intervals are shown graphically, and both the effect for the whole population (ATE) and the effect for the population of the treated (ATET) are represented.[34] Pairwise effects compare interventions that are the most similar, but that incrementally differ in their severity. For instance, the pairwise comparison of academic support and individual therapy compares interventions which are very similar and which target

[34] Regression tables with point estimates and exact confidence intervals are available upon request.



*Notes:* This figure depicts pairwise treatment effects for Special Education programs in St. Gallen. Each pair compares interventions that are the closest in degree of severity and inclusion. Each pairwise treatment effect is the effect of being assigned to the first program in comparison to the second program on one of the four outcomes presented in the panel headers. Both the treatment effect on the whole population (ATE) and on the population of the treated (ATET) are presented. "Ind. therapy" is the abbreviation for individual therapies, "Acad. support" for academic support, and "no SpEd" for receiving no program. Nuisance parameters are estimated using an ensemble learner that includes text representations presented in the "data" section. 95% confidence intervals are represented and are based on one sample $t$-test for the ATE and the ATET. Test results and wages are standardized with mean 0 and standard deviation 1. The baseline ("No SpEd") probability of unemployment benefit recipiency is 0.19, and 0.07 for disability insurance recipiency. *Source: SPS*.

Figure 2.5: Pairwise returns to Special Education programs according to their level of inclusion.

issues that are overlapping. The exception is counseling, which I compare to no treatment, as counseling does not happen in schools but with independent psychologists.

Results clearly show that returns to counseling are positive for academic performance. Students who receive counseling seem to fare better academically in comparison to those who do not receive any intervention but who exhibit similar difficulties and characteristics. This effect is four times the 0.1 standard deviation effect size criterion for successful interventions suggested by Bloom et al. (2006) and Schwartz, Hopkins, and Stiefel (2021). Academic support offers no benefits but does not harm either. Results suggest that individual therapies are more effective than tutoring to improve academic performance. This is due most likely to the fact that students in individual therapies work alone with a trained therapist who can address the roots of their learning difficulties (for instance, dyscalculia or speech problems). Finally, students in inclusive intervention fare worse than students in individual therapies. The main explanation for this difference is that inclusion is designed to address clusters of learning, psychological, behavioral and social problems, whereas individual therapies tackle one particular (learning) disability. Dealing with multi-faceted issues might render inclusion less effective in terms of academic achievement.

Long-term labor market effects are consistent with the effects on academic performance. As regards the probability of being unemployed and thus of benefiting from unemployment insurance, results show that counseling (-3 percentage points) and individual therapies (-10 percentage points) have a positive effect. Students benefiting from academic support are more likely to be unemployed than students with the same issues receiving no SpEd. A possible explanation for this negative labor integration effect is that these students needed support to succeed in school, support which is no longer provided once they enter the labor market. Individual therapies are more effective than tutoring in lowering unemployment probability (-10 p.p.), and are almost as effective as inclusion. Most pairwise effects indicate that intenser programs lead to lower probabilities of benefiting from disability insurance, with individual therapies showing the strongest reduction in disability insurance recipiency. The difference between inclusion and individual therapies in terms of disability insurance recipiency is small (0.8 percentage points for the ATE, 0 for the ATET). Finally, effects in wage returns remain small, as most pairwise effects are kept within the 0.1 standard deviation effect size criterion for successful SpEd interventions. Only individual therapies increase expected wage returns by almost 0.15 standard deviations in comparison to tutoring.

Figure A.3 in the Appendix shows that more severe interventions in the inclusive setting slightly increase the probability of taking the Stellwerk8 test. Although the

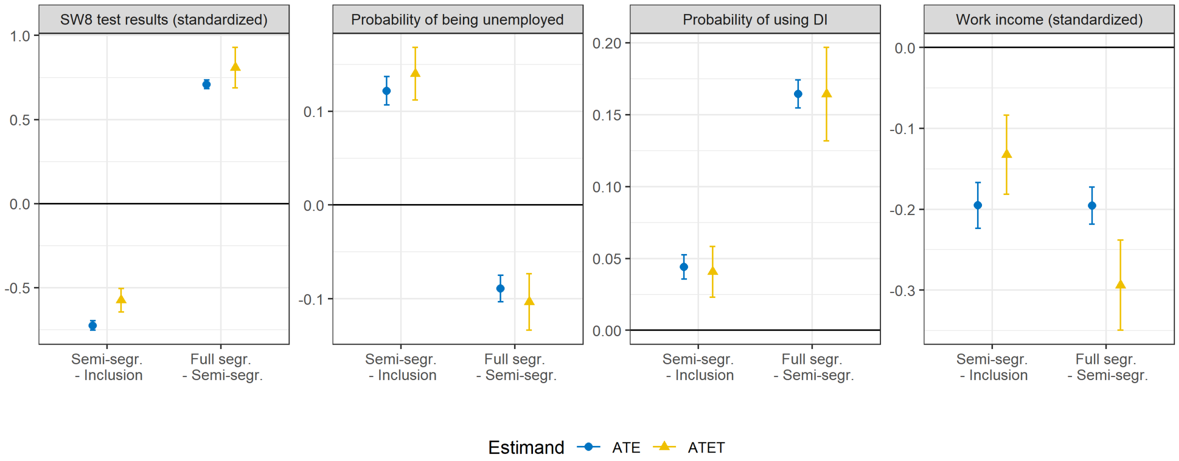
Stellwerk8 test does not indicate graduation *per se*, these results back the idea that SpEd interventions slightly increase the probability of attending high-stake tests close to the age of graduation, which brings nuance to the findings of Schwartz, Hopkins, and Stiefel (2021), McGee (2011) or Kirjavainen, Pulkkinen, and Jahnukainen (2016). Moreover, Figure A.3 shows that more serious interventions in inclusive settings are as effective as more benign interventions in ensuring that students with SEN take the test.

Finally, in most pairwise treatment effects, the ATE does not differ significantly from the ATET, which suggests that effects for the population of the treated are consistent with effects for the whole population. Noticeable differences between the ATE and the ATET persist for the pairwise comparisons that involve the "no treatment" category in the case of disability insurance. For these comparisons, the population of students who either receive counseling or academic support are more positively affected than the whole population by the interventions.

### 2.4.2 Returns to Special Education programs in segregated school settings

I now pay closer attention to returns to inclusion and segregation. I first compare inclusion and semi-segregation, which are two SpEd programs that are considered as close substitutes in St. Gallen. Second, I compare semi-segregation with full segregation. Results presented in Figure 2.6 speak in favor of inclusive measures when it comes to improving academic performance: students in inclusive settings perform on average 0.6 test score standard deviations better than students sent to semi-segregation. The most commonly given explanation for the success of inclusion in the literature is that mainstreaming enhances the performance of students with SEN due to a more stimulating and demanding environment (e.g., Cole, Waldron, and Majd, 2004; Daniel and King, 1997; Peetsma et al., 2001).

As regards labor participation, students sent to semi-segregation have a 10 percentage point higher probability to become unemployed than SEN students kept in the mainstream classroom. If students in inclusive settings have an average probability of being unemployed of around 11%, this probability reaches around 20% for students in semi-segregation. These findings confirm findings from Eckhart et al. (2011), who mention the lack of social network of segregated students as a plausible reason for



*Notes:* This figure depicts relevant pairwise treatment effects for Special Education programs in St. Gallen. Each pairwise treatment effect is the effect of being assigned to the first program in comparison to the second program on one of the four outcomes presented in the panel headers. Both the treatment effect on the whole population (ATE) and on the population of the treated (ATET) are presented. "Semi-segr." is the abbreviation for semi-segregation (segregation in small classes), and "Full segr." stands for full segregation (in special schools). Nuisance parameters are estimated using an ensemble learner that includes text representations presented in the "data" section. 95% confidence intervals are represented and are based on one sample $t$-test for the ATE and the ATET. Test results and wages are standardized with mean 0 and standard deviation 1. The baseline ("Inclusion") probability of unemployment benefit recipiency is 0.19, and 0.04 for disability insurance recipiency. *Source: SPS.*

Figure 2.6: Pairwise returns to segregated SpEd programs.

lower employment. Moreover, these results might be explained by the fact that semi-segregation is attached to a signaling penalty, i.e. semi-segregation results in an irregular degree that considerably reduces access to regular VET programs. This is even more striking as lower employment by semi-segregated students comes exclusively from unemployment insurance and not from disability insurance. Estimates show a significant wage gap: students placed in semi-segregated settings earn on average 0.15 standard deviations less than students placed in the main classroom.

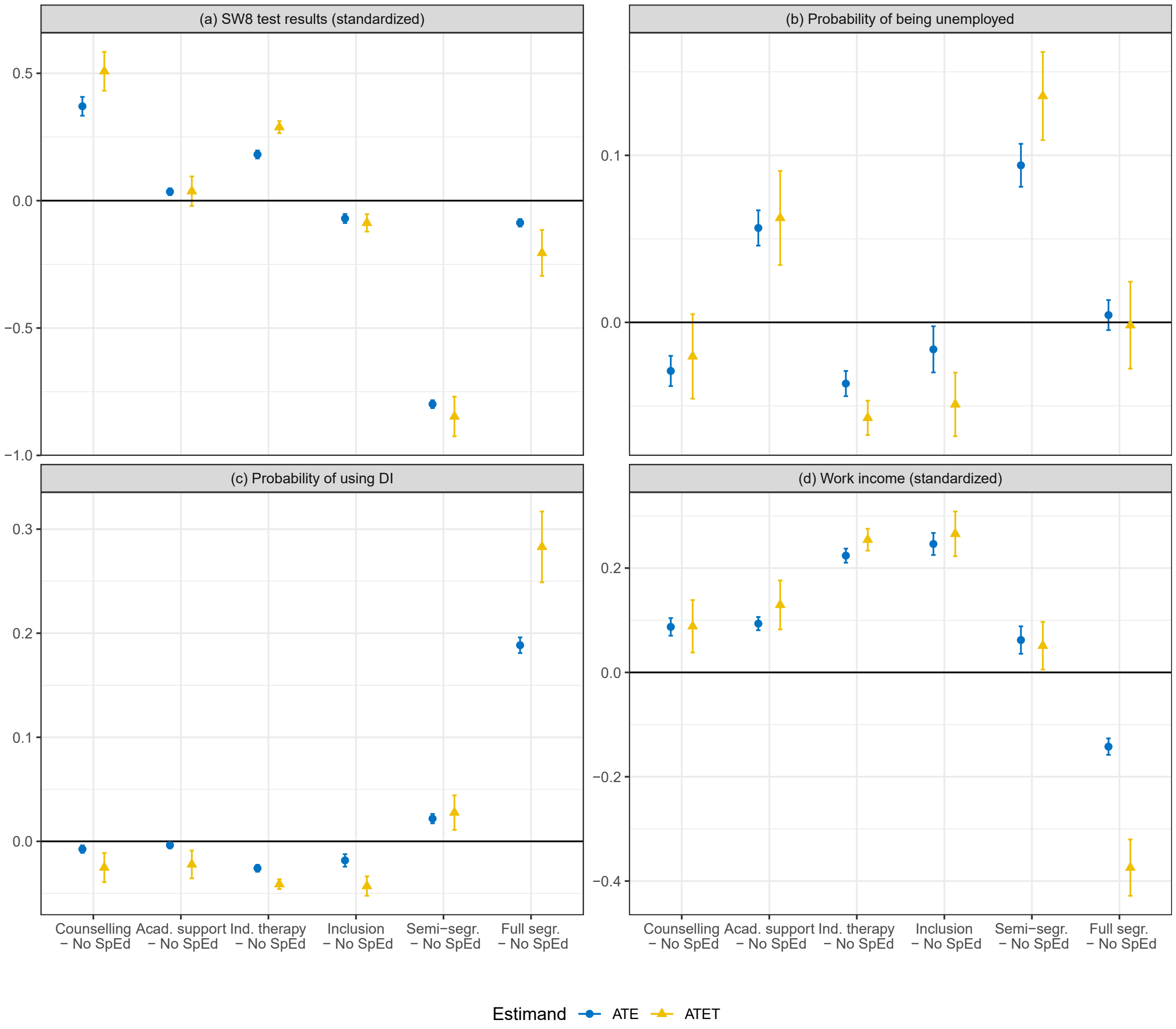
Returns to full segregation in comparison to semi-segregation are grim: the seemingly positive returns to full segregation in terms of academic performance are mostly due to selection into test participation (see Figure A.3). Since, in the Canton of St. Gallen, only students in segregated schooling environments are allowed to opt out of the mandatory test, SEN students in full segregation are between 20 to 30 percentage points less likely to take the SW8 test than students in semi-segregation. Contrary to the case of semi-segregation, results for full segregation correspond to the findings of McGee (2011) and Kirjavainen, Pulkkinen, and Jahnukainen (2016), who found a lower graduation rate for students in SpEd. This emphasizes the importance of making a distinction between semi-segregation and full segregation when assessing the effects of segregation in general.

SEN students assigned to segregated schooling have a significantly higher probability of benefiting from disability insurance, but not of becoming unemployed. Interestingly, the channels of (lack of) labor market participation are different for semi-segregated and fully segregated students: whereas students in small classes are more likely to be unemployed, students in special schools are likely to end up on disability insurance. This is not surprising in light of the information extracted from the psychological records. Records strongly suggests that students in full segregation are students with a higher probability of suffering from motor disabilities, developmental problems as well as speech and language problems (see Figure A.4). Students with SEN are less likely to be sent to full segregation for learning disabilities. As mentioned in the identification section, results about full segregation should be interpreted with caution, as overlap is difficult to obtain for students in full segregation.

### 2.4.3 Returns to Special Education interventions against no intervention

Most of the literature evaluating SpEd programs focuses on returns to SpEd as a single intervention, without distinguishing between SpEd programs. My results clearly show that SpEd cannot be considered as a single intervention, and that each program brings different returns. In order to compare my results to results presented in the existing literature, I now compare the effect of each program with receiving no program at all. Even though all effects take into account observed confounding from covariates and text, it is clear that the ATE and ATET between receiving an intervention and not receiving one becomes more and more difficult to identify as interventions become more severe (lack of overlap). I therefore remain cautious when interpreting the effect of receiving no intervention with receiving more intensive interventions such as semi-segregation.

Figure 2.7 presents main results and shows intervention effects from the least intensive interventions (left) to the most intensive interventions (right). I represent again the first two pairwise effects for sake of comparison. Counseling and individual therapies have positive academic returns (their effects exceed the threshold of 0.1 standard deviations in test score). Academic support and inclusion bring almost zero returns



*Notes:* This figure depicts relevant pairwise treatment effects for Special Education programs in St. Gallen. Each pairwise treatment effect is the effect of being assigned to the first program in comparison to being assigned to no program on one of the four outcomes presented in the panel headers. Both the treatment effect on the whole population (ATE) and on the population of the treated (ATET) are presented. “Ind. therapy” is the abbreviation for individual therapies, “Acad. support” for academic support, “no SpEd” for receiving no program, “Semi-segr.” for semi-segregation (segregation in small classes), and “Full segr.” for full segregation (in special schools). Nuisance parameters are estimated using an ensemble learner that includes text representations presented in the “data” section. 95% confidence intervals are represented and are based on one sample *t*-test for the ATE and the ATET. *Source: SPS*.

Figure 2.7: Pairwise returns to Special Education vs. No Special Education.

(for inclusion, I measure -0.0878 for the ATE and -0.121 for the ATET). Students with SEN in these four programs catch up on the labor market: they are less likely to be unemployed (Figure 2.7b.), are less likely to end up on disability insurance (Figure 2.7c.), and earn as much or even more than students having received no SpEd (Figure 2.7d.). Only academic support does not improve chances of labor market integration.

Effects of segregation are generally negative. Segregated interventions have negative academic returns (the negative effects of full segregation are due to attrition into test taking), and generate negative labor market outcomes. Students with SEN in semi-segregated settings have a higher probability of becoming unemployed (around 10 percentage points), but not of receiving DI. However, returns to semi-segregation in terms of wages are similar to receiving no SpEd at all. In contrast, students with SEN assigned to segregated schooling have a significantly higher probability of benefiting from disability insurance, but not of becoming unemployed.

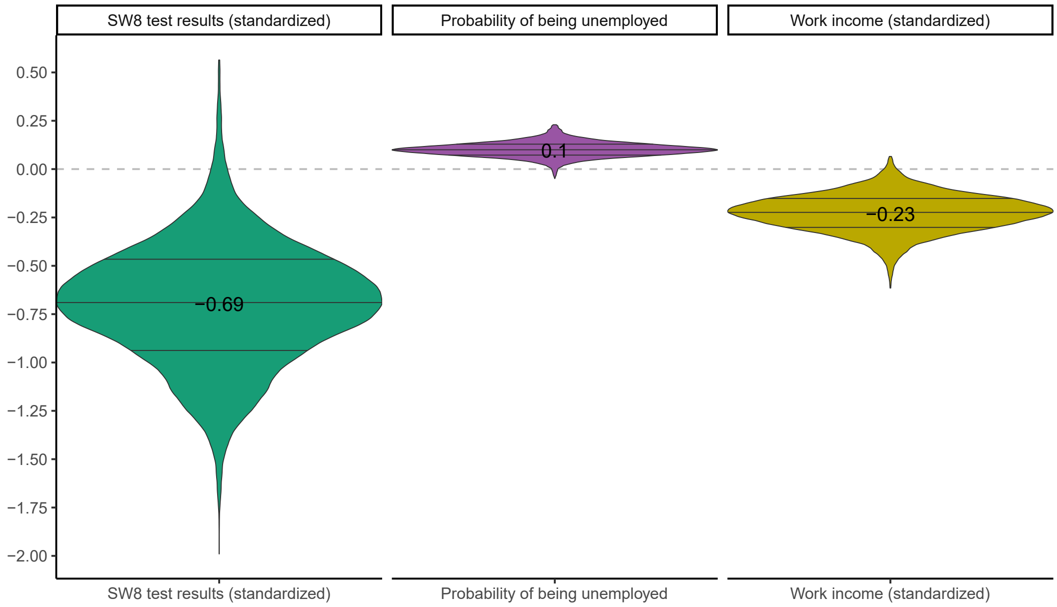
### 2.4.4 Inclusion and semi-segregation: who benefits from segregation?

Is inclusion always preferable to segregation for students with SEN? Despite the (political) decision to implement inclusive SpEd rather than segregated SpEd, a nuanced analysis about which students might still benefit from segregation is lacking. In this section, I explore whether average effects of inclusion compared to semi-segregation hides treatment effect heterogeneity, and whether semi-segregation might be beneficial for some students.

I first investigate Individual Average Treatment Effects (IATEs) for the effect of inclusion in comparison to semi-segregation. IATEs give the treatment effects at the most granular level and are useful to identify students with SEN who benefit the most from each treatment assignment. Figure 2.8 reports the smoothed distribution of IATEs per outcome with the light horizontal bars depicting the first and fifth quintiles of the IATE distribution.[35] The distributions of IATEs are quite spread out, indicating large heterogeneities in responses to semi-segregation in comparison to inclusion across students with SEN. This is specially salient for IATEs in academic performance. For all outcomes, some students are shown to be indifferent between semi-segregation and inclusion, or even to benefit from semi-segregation.

To have an idea of which individual characteristics are the most predictive of treatment effect size, I perform a classification analysis in the spirit of Chernozhukov,

[35] Note that some predicted IATEs have very large values, especially for the test scores. This is due to the fact that the DR-learner is weighted by the inverse of the propensity scores, which do not sum to one in finite samples. Although I am not concerned with extreme values in this case, I computed the normalized DR-learner of Knaus (2021), which mitigates this problem. Results are very similar, and are available upon request.



*Notes:* This graph represents the smoothed distribution of IATEs for the four outcomes of interest. The IATEs were predicted out-of-sample using the DR-learner presented in Equation (2.6). The average effect is reported in the figure. Light quintile bars represent the median and the 1st. and 5th. quintiles of the IATE distribution.

Figure 2.8: Distributions of IATEs for segregation vs. inclusion.

Fernández-Val, and Luo (2018). I group the predicted IATEs into quintiles and compare the standardized means difference (SMD) of covariates for students with the 20% highest IATEs (fifth quintile) and students with the 20% lowest IATEs (first quintile). For the probability of unemployment, the first and fifth quintiles are reversed in the table, as students in the fifth quintile suffer the most from semi-segregation. SMDs that are larger than 0.2 are considered to be large (Rosenbaum and Rubin, 1985). I report covariates for which the standardized mean difference is higher than 0.2 in at least one of the treatment effects. Note that I do not report results for disability insurance, as average effects are almost zero.

Table 2.3 shows the characteristics of students in the lower and higher tails of the IATE distributions according to main covariates. Students in the highest IATE quintile for academic performance are more likely to be nonnative students referred for social and emotional issues. Students in need of psychological support are also more likely to benefit from semi-segregation for academic performance. This particular population of students with SEN is also more likely to benefit from inclusion in terms of employment and wages. Finally, results clearly show that the age at referral is important for labor market outcomes: students that are referred earlier to the SPS clearly

| | **Test scores** | | | **Unemployment Probability** | | | **Wage** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Quint. I. | Quint. V. | SMD | Quint. V. | Quint. I. | SMD | Quint I. | Quint. V. | SMD |
| **Main covariates** | | | | | | | | | |
| Female | 0.45 | 0.38 | 0.14 | 0.41 | 0.39 | 0.04 | 0.48 | 0.35 | **0.28** |
| Nonnative | 0.07 | 0.34 | **0.73** | 0.06 | 0.22 | **0.48** | 0.04 | 0.21 | **0.55** |
| IQ | 93.41 | 97.20 | **0.34** | 94.66 | 94.82 | 0.01 | 95.82 | 93.66 | 0.18 |
| Age referral | 8.00 | 8.39 | 0.19 | 10.06 | 8.61 | **0.63** | 10.06 | 8.40 | **0.75** |
| Referral: social/emotional problems | 0.13 | 0.37 | **0.56** | 0.13 | 0.28 | **0.39** | 0.13 | 0.27 | **0.34** |
| Referral: performance/learning problems | 0.87 | 0.90 | 0.09 | 0.91 | 0.79 | **0.35** | 0.85 | 0.84 | 0.03 |
| Referral: conflict with teacher | 0.02 | 0.05 | **0.21** | 0.01 | 0.04 | 0.141 | 0.01 | 0.03 | 0.11 |
| Need psychological treatment | 0.13 | 0.26 | **0.31** | 0.14 | 0.24 | **0.27** | 0.11 | 0.25 | **0.35** |
| Nonnative × Female | 0.04 | 0.13 | **0.34** | 0.03 | 0.08 | **0.22** | 0.02 | 0.08 | **0.28** |
| Nonnative × Social/emotional | 0.01 | 0.11 | **0.43** | 0.01 | 0.06 | **0.31** | 0.00 | 0.06 | **0.33** |

Table 2.3: Classification analysis of IATEs for semi-segregation vs inclusion

*Notes:* This table shows the mean of each covariate and the standardized mean differences (SMD) across both SE programs between the fifth and the first quintile of the respective estimated IATE distribution. The first quintile refers to the most negative effects, whereas the fifth quintile to the most positive effects. For the probability of unemployment, the first and fifth quintiles are reversed (as students in the fifth quintiles suffer the most from semi-segregation). For two SE programs $w$ and $w'$, SMDs are computed as $\frac{\bar{x}_w - \bar{x}_{w'}}{\sqrt{\frac{s_w^2+s_{w'}^2}{2}}}$, where $\bar{x}_w$ is the mean of the covariate in treatment group $w$ and $s_w^2$ is the sample variance of covariate in treatment group $w$. SMDs higher than 0.2 are depicted in bold.

better benefit from segregation in terms of labor market integration. It is important to note that gender alone is not a predictive characteristic of different individual effects, the exception being that female students can expect worse wage outcomes as a result of semi-segregation.

### Group heterogeneity in the effect of semi-segregation: the disruption hypothesis

An argument in favor of semi-segregation is that it attenuates disruption in the main classroom by removing "disruptive" students from the mainstream environment. However, the question whether semi-segregation is more beneficial than inclusion for "disruptive" peers has not yet been answered in the literature. Disruptive students are students who disturb their classmates and need additional teacher time and attention (see Lazear, 2001; Carrell, Hoekstra, and Kuka, 2018). They might benefit from a segregated environment, which offers them increased teacher time and the right monitoring to focus on academic tasks. In this section, I explore whether returns to inclusion and semi-segregation systematically differ along pretreatment characteristics that potentially reveal disruptive behaviors: gender, nonnative speaking, whether the student has been referred for behavioral problems, and interaction between these groups. Disruptive behaviors are known to be prevalent in male students (Bertrand and Pan,

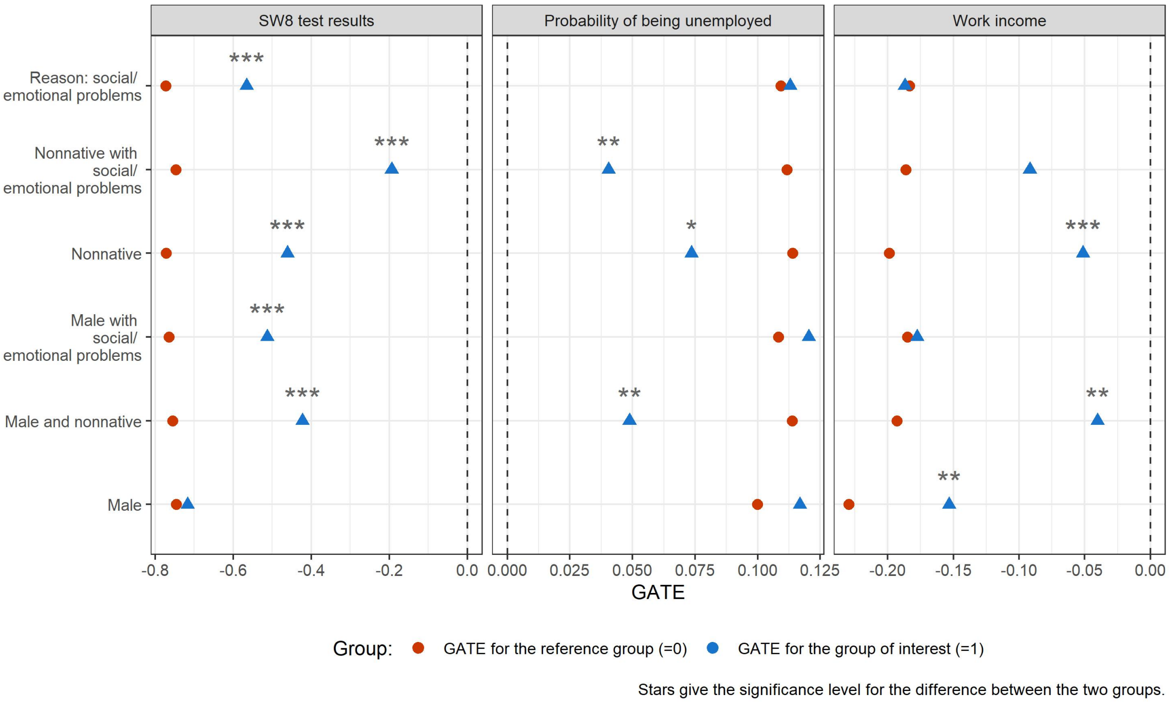


Figure 2.9: GATEs for semi-segregation vs inclusion

*Notes:* This graph presents estimated GATEs for the treatment effect of semi-segregation vs. inclusion. The GATEs are represented on the $x$ axis. Red dots indicate the treatment effect for the reference category (those students who do not belong to the category of interest), and blue dots indicate the treatment effect for the category of interest. The stars indicate the statistical significance of the difference between the two groups. For instance, the first row of the first column shows that the treatment effect of semi-segregation vs. inclusion is -0.75 test score standard deviations for students without social or emotional problems. The treatment effect is -0.55 for students with social or emotional problems. The difference in treatment effect between both groups is statistically significant ($p < 0.01$).

2013; Lavy and Schlosser, 2011b), students with behavioral problems (Fletcher, 2009), or nonnative speakers (Diette and Uwaifo Oyelere, 2014; Cho, 2012).

Findings about the "disruptiveness" hypothesis are presented in Figure 2.9. Each row of the figure gives the results of a regression where the DR score for the ATE is regressed on the group dummy. The GATEs are represented on the $x$ axis. Red dots indicate the treatment effect for the reference category (those students who do not belong to the category of interest), and blue dots indicate the treatment effect for the category of interest. The stars indicate the statistical significance of the difference between the two groups. For instance, the first row of the first column shows that the treatment effect of semi-segregation vs. inclusion is -0.75 test score standard deviations for students without social or emotional problems. The treatment effect is around -

0.55 for students with social or emotional problems. The difference in treatment effect between both groups is statistically significant ($p < 0.01$).

From this analysis, two main conclusions can be drawn. First, heterogeneity in effects along disruptive characteristics are important for school performance, and also for long-term outcomes. Major effect differences between inclusion and semi-segregation can be found for male nonnative students with social or emotional problems. Second, results of the GATE analysis clearly show that "disruptive" students tend to benefit more from semi-segregation than non-disruptive students. However, my analysis does not show that "disruptive" students would perform better in semi-segregation settings: they would still be better off in inclusive settings. In particular, while semi-segregation negatively impacts SEN students on average, three particular groups of SEN students seem to have systematically higher GATEs: nonnative speakers, students with social and emotional problems, and male students. Any subgroups of students among these three groups exhibit GATEs that are higher than the ATEs. For instance, the effect on test scores of semi-segregation in comparison to inclusion is 0.3 standard deviations higher for nonnative speakers than for native speakers. Nonnative speakers are also less likely to be unemployed when segregated than native speakers, and they expect higher wages. When segregated, they expect a wage premium of .15 wage standard deviations higher than for natives, making semi-segregation as good as inclusion in terms of expected wages. The subgroup that would see the smallest difference between inclusion and semi-segregation are nonnative students with emotional or behavioral problems, who would perform only .19 test score standard deviations less in segregation than in inclusion (and almost 0.6 standard deviations better than other students in semi-segregation). All in all, even though inclusion remains on average better for students with SEN, even for those with "disruptive" characteristics, my results show that "disruptive" students are the students who benefit the most from semi-segregation.

**Do the effects of semi-segregation vary with school characteristics?**

School characteristics, especially the school socio-economic composition (SES score), could potentially drive the effect of inclusive and semi-segregated programs. To investigate this mechanism, I explore treatment heterogeneity in school characteristics for inclusion and semi-segregation. To estimate CATEs with respect to school characteristics, I regress the continuous school characteristics on the DR score for the ATE following Zimmert and Lechner (2019) as explained in Section 2.3.3. I present results for the CATE for academic test score and for the probability to be unemployed in Figure A.4 and Figure A.6. I do not report per-student spending in my heterogeneity analysis given that it has been measured only once in 2017 and would not reflect schools' change of SpEd policies across the years.

Results show an interesting tendency in the effects of inclusion in comparison to semi-segregation. Negative effects in school performance get closer to 0 as school size increases, and also as the schools' SES score and the share of nonnative students increase. For instance, the negative effects of semi-segregation become 0.2 standard deviations smaller when the school has a high SES score (meaning that the school's population is of a lower average SES status). The magnitude of this effect is rather substantial, and would indicate that students in schools that have a less homogeneous population (mostly in urban schools) and with more lower SES students are not harmed as much if they are segregated. The same picture is reflected in the effects for the probability to be unemployed: the negative effects of semi-segregation become almost zero in schools with a high number of foreign language speaking students, and with a higher SES score.

Thus, the negative effects of semi-segregation are shown to slightly fade out for schools that are more socially and economically diverse. One potential explanation for this phenomenon is that the average level of disruptiveness in schools with more diversity would make semi-segregation comparatively more effective for disruptive students. Another possible explanation is that inclusion works better when teachers are less overwhelmed, i.e. in more homogeneous classrooms. These findings indirectly corroborate teachers' worries that inclusion is difficult to implement when the population of students is already difficult to deal with in the main classroom.[36]

In conclusion, my analysis shows that semi-segregation programs are less effective than inclusion in terms of academic performance and labor market integration for (almost) all students with SEN. However, students who exhibit "disruptive" characteristics, i.e. male students, students with social and emotional problems, and nonnative speakers, are the students who benefit most from semi-segregation. Importantly, the success of inclusion or semi-segregation does not depend on IQ, gender or the preva-

[36] Such arguments have been expressed by teachers: inclusion is a factor of teachers' overload, incompatibility between non-SEN and included SEN students, and a lack of efficiency for SEN students. Newspaper articles from the *Tagblatt* newspaper in St. Gallen regularly refer to the problem. For instance *"How inclusion divide teachers"*, *"Special education needs resources"*, *"Include instead of segregate"*, *"Teaching the same to students who are not the same"*, *"The integrative model puts teachers to their limits"*.

lence of learning disabilities: this is especially true for school performance, but also extends to labor market integration. Finally, the schools' average SES status and population of nonnative students has non-negligible influence on the effect of inclusion and semi-segregation SpEd programs.

### 2.4.5 Further analyses and robustness checks

I conduct a battery of robustness tests in Appendix A.2. In Appendix A.2.1, I account for potential overlap and lack of common support problems in the generalized propensity score distribution and implement overlap-weighted average treatment effects (Li, Morgan, and Zaslavsky, 2018; Li and Li, 2019) as well as different trimming schemes (see Crump et al., 2009; Stürmer et al., 2010). I find that point estimates do not vary much when extreme weights are trimmed, and become less precisely estimated.

In Appendix A.2.2, I investigate how sensitive my estimates are to the inclusion of text covariates in order to see how much confounding my text variables are able to remove. I find that estimates based on both covariates and text information are on average 29% smaller than estimates that do not leverage the text information. Furthermore, I account for the possibility that my text representations capture the psychologist's biases (towards a certain treatment or a certain writing style) rather than information on the student. I show that my estimates remain consistent with my main findings when I strip out psychologists' effects from the text.

In Appendix A.2.3, I tackle the problem of potential selective attrition in the measured outcomes. I conduct an attrition analysis by showing the results for the cohorts that are in both the SW8 and the SSA subsamples. Results of this check are in line with main results.

Finally, even though I observe the assignment process in its entirety through written reports, some factors influencing SpEd placements might still remain unobserved. I address this issue by leveraging the within year across school variation in supply of SpEd as an instrument to compare the outcomes of SEN students in inclusive settings with similar peers in semi-segregated settings (in the spirit of Keslair, Maurin, and McNally, 2012). In Appendix A.2.4, I present Local Average Treatment Effects of inclusion on students who would have been segregated, had they lived in a school that implemented semi-segregated SpEd programs. LATE results corroborate my main findings.

## 2.5 Policy learning: optimal SpEd placement for inclusion and semi-segregation

How would a school psychologist or a policy maker assign SEN students to the program that corresponds the best to their particular characteristics? Getting insights from the literature on statistical treatment rules (e.g., Kitagawa and Tetenov, 2018; Manski, 2004), I perform optimal treatment allocations simulations based on tree-search based algorithms (Athey and Wager, 2021; Zhou, Athey, and Wager, 2018). I leverage my rich set of covariates as well as the information retrieved from the psychological records to look at whether policy makers might be able to better tailor policies on the basis of observed individual characteristics. I focus on allocation to inclusion and semi-segregation, two programs that are used as quasi-substitutes in St. Gallen.

Let $\pi(Z_i)$ be a policy rule that leverages a relatively small number of observed individual characteristics of interest $Z_i$ to assign individuals to a treatment $d$. The number of characteristics in $Z_i$ is small because it contains variables that can be easily interpreted and used by a policy maker. For each policy rule $\pi(Z_i)$ among all candidate policy rules $\Pi$, a policy value summarizes the estimated population potential outcome attained under the policy. The policy value is the average of the individual APOs (defined by Equation (2.6)) under the policy rule $\hat{Q}(\pi) = \frac{1}{N} \sum_{d=1}^{D} \sum_{i=1}^{N} \underline{1}(\pi(Z_i) = d)\hat{\Gamma}_i^d$. The goal of the policy maker is to find the optimal policy, i.e. the particular policy rule among all possible policy rules that maximizes the policy value. This means finding the policy rule $\hat{\pi}^*$ that maximizes the policy value function $\hat{\pi}^* = \arg\max_{\pi \in \Pi} \hat{Q}(\pi)$ among all candidate policies. Note that the goal of this exercise is not to discriminate students on the basis of their characteristics, but to guide policy makers on how they could potentially improve policies by leveraging observed characteristics.

For instance, a policy maker could propose three candidate policy rules for optimal placements: (1) assign all SEN students to inclusion; (2) assign all SEN students to semi-segregation; (3) "split" students along their nonnative status, i.e. assign all nonnative speakers to semi-segregation and all other students to inclusion. The policy value of the first rule is the average of all APOs under inclusion, and similarly for the second rule under semi-segregation. For the third rule, the policy value would be the average over the APO under semi-segregation for all nonnative students and over the APO under inclusion for all native students. The policy maker would then choose the policy with the highest overall policy value.

To find the optimal policy, I compute the policy tree algorithm with fixed depth based on double machine learning of Zhou, Athey, and Wager (2018).[37] Intuitively, trees (policy rules) split the sample along variables in $Z_i$ many times, and the split that gives the maximal policy value is given as the optimal policy. I experiment with policy trees of depth 2 and 3 (the depth indicates the number of times leaves are split) with two different sets of student attributes $Z$, i.e. the baseline covariates and covariates extracted from the diagnosis (dictionary approach). I compute optimal policy allocations for inclusion or semi-segregation on the subsample of students either sent to inclusion or segregation. As outcomes, I look at test scores and labor-market integration (probability of employment). I subsequently link optimal policy allocations to actual and counterfactual policy costs, which are estimated by taking average costs per student per year of each SpEd placement given by the Canton of St. Gallen.[38] I then compare changes in costs and changes in policy value for each optimal policy.

Panel A of Table 2.4 compares the actual observed percentage of students assigned to inclusion and to semi-segregation with the allocation of students resulting from optimal reallocation. Around 70% of the students sent either to inclusion or semi-segregation who took the SW8 test were actually assigned to inclusion, and 53% of students who were registered in the SSA dataset were assigned to inclusion. For academic performance, the policy value of the actual implemented policy is -0.46, and it is 0.67 for labor market integration (which gives the average employment probability under the implemented policy). I propose four reallocation policies for each outcome. All proposed policies dramatically improve the policy value by reallocating the majority of students to inclusion rather than segregation. All proposed policies would reduce overall costs, but not dramatically so (at around 94% of actual realized cost). In general, proposed assignment schemes are very similar in terms of improved outcomes and reduced costs.[39]

[37] The algorithm finds the optimal policy such that it minimizes the regret function, i.e. the difference between the true and the estimated optimal policy value. In general, see algorithm 1 in Zhou, Athey, and Wager (2018).

[38] For costs, I use the following estimates obtained from *SG-Volksschulgesetznachtrag 2013*. A student in mainstreamed environment costs between 15’000 and 20’000 CHF (approximately 16’500 to 22’000 USD) on average per year, depending on the school and the grade. I take the highest estimate, namely 20’000 CHF. A student in semi-segregation costs on average 24’500 CHF (27’000 USD) per year. Individual, hour-long therapy SpEd programs costs on average 5000 CHF (5500 USD) per year. Schooling in full segregation settings costs between 39’000 and 260’000 CHF, on average between 70’000 to 80’000 CHF (77’000 USD to 88’000 USD) per student per year.

[39] One could think that the remaining students sent to segregation are observations with extreme weights in their doubly robust score. I computed the same policy classification after trimming, and the policy rules would consistently send the same amount of students to segregation.

**Panel A**: Allocation to program in percent and potential cost reduction

| | % Students sent to inclusion | % Students sent to semi-segregation | Policy value | Costs per year (in mm CHF) | Percent of actual costs |
|---|---|---|---|---|---|
| ***Test scores. N = 2988*** | | | | | |
| *Actual allocation* | 0.69 | 0.31 | -0.46 | 63,925 | 1 |
| Depth 2 and baseline variables | 0.96 | 0.04 | -0.29 | 60,271 | 0.94 |
| Depth 3 and baseline variables | 0.96 | 0.04 | -0.27 | 60,235 | 0.94 |
| Depth 2 and diagnosis variables | 0.96 | 0.04 | -0.29 | 59,762 | 0.93 |
| Depth 3 and diagnosis variables | 0.97 | 0.03 | -0.27 | 60,145 | 0.94 |
| ***Probability of employment. N = 2939*** | | | | | |
| *Actual allocation* | 0.53 | 0.47 | 0.67 | 64,958 | 1 |
| Depth 2 and baseline variables | 0.83 | 0.17 | 0.79 | 61,007 | 0.94 |
| Depth 3 and baseline variables | 0.87 | 0.13 | 0.80 | 60,490 | 0.93 |
| Depth 2 and diagnosis variables | 0.87 | 0.13 | 0.79 | 60,485 | 0.93 |
| Depth 3 and diagnosis variables | 0.85 | 0.15 | 0.81 | 60,796 | 0.94 |

*Notes:* This table shows the treatment assignment from four different policies. The depth indicates the depth of the policy trees. The baseline variables are individual covariates excluding variables from text, and diagnosis variables are individual covariates together with covariates from the diagnoses extracted from the text (dictionary approach). The policy value is the average APO under each policy. Total cost estimates and potential cost reduction from the implemented policies are computed.

**Panel B**: Cross-validated difference between optimal policy value and different policies

| | All inclusion | All semi-segregation | Assigned policy |
|---|---|---|---|
| ***Test scores. N = 2988*** | | | |
| Depth 2 and baseline variables | -0.010$^{**}$ | 0.651$^{***}$ | 0.484$^{***}$ |
| | (0.004) | (0.036) | (0.029) |
| Depth 3 and baseline variables | -0.009$^{*}$ | 0.652$^{***}$ | 0.485$^{***}$ |
| | (0.004) | (0.036) | (0.029) |
| Depth 2 and diagnosis variables | -0.009$^{**}$ | 0.652$^{***}$ | 0.485$^{***}$ |
| | (0.004) | (0.036) | (0.029) |
| Depth 3 and diagnosis variables | -0.015$^{*}$ | 0.646$^{***}$ | 0.478$^{***}$ |
| | (0.009) | (0.035) | (0.029) |
| ***Probability of no unemployment. N = 2939*** | | | |
| Depth 2 and baseline variables | -0.027$^{***}$ | 0.110$^{***}$ | 0.016 |
| | (0.008) | (0.037) | (0.028) |
| Depth 3 and baseline variables | -0.031$^{**}$ | 0.106$^{***}$ | 0.012 |
| | (0.013) | (0.035) | (0.029) |
| Depth 2 and diagnosis variables | -0.021$^{**}$ | 0.116$^{***}$ | 0.022 |
| | (0.009) | (0.037) | (0.028) |
| Depth 3 and diagnosis variables | -0.043$^{***}$ | 0.094$^{***}$ | -0.000 |
| | (0.014) | (0.035) | (0.030) |

*Notes:* This table displays validation tests for policy trees. 10-fold cross-validation is used. Optimal policies are compared to either sending everyone to inclusion, to sending everyone to semi-segregation, or to the (already implemented) observed policy. For each policy, the average difference between the APO under the optimal policy and the APO under one of the three alternative policies is computed. Inference is done with a one sample $t-$test on the difference. Standard deviations of one-sample $t$-test in parenthesis. $^{***}$: p <0.01, $^{**}$: p < 0.05, $^{*}$: p < 0.1.

Table 2.4: Optimal policies for inclusion and semi-segregation

What are the rules that allow better allocation of students in terms of better school performance? Figure A.8 represents the tree policies for improved test scores. Interestingly, a simple policy tree of depth 2 would automatically assign all nonnative students with emotional or social issues to semi-segregation, as well as all gifted students (with IQ higher than 125) without emotional or social issues to semi-segregation. All other students with SEN would be assigned to inclusive settings. This makes sense since students with social problems are likely to be disruptive and might benefit from segregation, and since gifted students might benefit from "pull-out" programs. This policy would outperform the actual assignment by gaining an average of around 0.2 test score standard deviations.[40] The policy tree of depth 3 with baseline characteristics confirms the importance of issues in emotional or social behaviors, problems with test performance, and IQ score as important variables for optimal policies. Policies using diagnoses are very similar to policies based solely on main covariates, which highlights the fact that diagnoses extracted from text are not very predictive of policy improvement.

Sending more students to inclusive settings has benefits for integration on the labor market. The second part of Panel A in Table 2.4 shows that by sending less students to semi-segregation, the average employment probability can be increased by around 20 percentage points for quasi-similar policy costs. Interestingly, proposed policies that target better labor-market integration send a larger share of students to semi-segregation. Figure A.10 shows that semi-segregation is the most helpful for students with IQ scores lower than average, without performance or learning problems. In general, IQ seems to be the most predictive variable of success of semi-segregation with respect to labor-market integration. Students who were not given an IQ test would also be sent to semi-segregation. From the trees with diagnoses, we learn that children with ADHD would be sent to semi-segregation. The policy tree of depth 3 that leverages the information on ADHD is the most effective policy in terms of labor market integration.

To test whether the decision trees are stable, I conduct validation tests inspired by Zhou, Athey, and Wager (2018) and Knaus (2021). I test whether the proposed policies perform better than either sending everyone to inclusion, sending everyone to semi-segregation, or implementing the (already implemented) observed policy (the

[40] Note that the variable "IQ score" is actually the interaction between the actual score and whether the IQ test has been administered, thus taking 0 when no IQ score exists. I perform a similar analysis with the subsample of students for whom the IQ is observed.

"Null-hypothesis" policies). To do this, I use 10-fold cross-validation, i.e. I train the policy tree on the training subsample and use the tree to predict assignment on the left-out fold. I then compute the difference between the APO under the optimal policy and the APO under one of the three alternative policies, and compute a one sample $t-$test to assess whether the difference in means is significantly different than zero. Panel B in Table 2.4 shows that all optimal reallocations outperform sending all students to semi-segregation for both outcomes. For academic performance, all policies outperform the implemented policies, but perform as well or even slightly worse than sending all students to inclusion. For labor-market integration, sending all students to inclusive settings marginally outperforms the optimal policies; however, all proposed policies fail to significantly improve on realized therapy assignments. This indicates that actual placement by school psychologists already leverages the available information in a relevant way, but that this placement could be improved to increase academic performance.

I implicitly assumed in my optimal allocation exercise that non-SEN students receive a welfare weight of 0. In reality, the reallocation of students with SEN from semi-segregation to the main classroom induce spillover effects that could negatively impact mainstreamed SEN and non-SEN students (as shown by Balestra, Eugster, and Liebert, forthcoming; Rangvid, 2019). To measure overall welfare functions, I integrate spillover effects of reassigning students with SEN to the main classroom in my optimal policy rule. I proceed as follows: I merge my dataset with the data containing all students without SEN from the Canton of St. Gallen. I then estimate the peer effect of reassigning students with SEN to mainstream classrooms on the academic performance of their classmates (for classmates with SEN and for classmates without SEN). As students with SEN are randomly allocated to classrooms in St. Gallen, I am able to estimate causal spillover functions with flexible estimation procedures (for more information on this quasi-experimental setting, see Balestra, Sallin, and Wolter, forthcoming).[41] The classroom spillover function is presented in Figure A.12 of the Appendix. The function is estimated for students with SEN (in brown) and for students without SEN (in blue). I find that the shape of the spillover function is monotonically decreasing, meaning that including an additional peer with SEN has negative effects on mainstreamed peers with SEN and on mainstreamed peers without SEN. The negative effect worsens with more students with SEN in the classroom.

[41] I estimate spillovers with machine learning algorithms and an ensemble learner. I use clustered cross-validation to estimate functions at the classroom level.

Based on my flexible estimation of the spillover function, and with a simplistic back-of-the-envelope utility computation, I compute the average reallocation effects induced by the optimal policy for the whole student population. I base my computations on the following statistics: students with SEN make up 25% of the mainstreamed students population, mainstream classrooms have on average 19.17 students, and the data contains 2723 classrooms. As presented in Panel A of Table 2.4, optimal policies reallocate 807 students from semi-segregation to inclusion for a policy gain of 0.17. The reallocation of 807 students from semi-segregation to inclusion means that there are 0.3 additional students with SEN per classroom, thus increasing the average proportion of students with SEN per classroom from 0.25 to 0.266 (a 0.016 increase). The effect of reallocation for the reallocated students is therefore an average gain of $0.016 * .17 = 0.003$ in policy value per student in the whole population. For the population of mainstreamed students, the spillover functions in Figure A.12 (and tables available on request) show that the increase in the proportion of peers with SEN in the main classroom from 0.25 to 0.266 generates a loss of expected test score standard deviation of -0.04 for mainstreamed students with SEN, and a loss of -0.03 test score standard deviations for mainstreamed students without SEN. This means that the policy loss is on average around $0.25 * (-0.04) + 0.75 * (-0.03) = -0.032$ per student. Combining both effects (and ignoring the very small shift in group proportions after reallocation), the average loss in policy value after reallocation is around $-0.03$ standard deviations of test scores per student.

All in all, this optimal policy exercise exercise delivers valuable insights into improved allocation of SEN students to semi-segregation and inclusion. It strengthens the idea that inclusion works well, as none of my suggested reallocation policies perform better than allocating all students to inclusion. There also seems to be a trade-off between short-term, academic benefits, and longer-term benefits: from the perspective of labor-market integration, it seems that a higher share of semi-segregated students is beneficial, whereas it is not from an academic performance perspective. Finally, it highlights the idea that less segregation is desirable and could be reached without imposing too much harm on students in mainstream classrooms: the individual gains for the newly reallocated students are large and by far offset the individual loss for mainstreamed students, even though the social reallocation effect is negative by $-0.03$ standard deviations of test scores per student. In light of the results from the heterogeneity analysis and from the reallocation exercise, it might seem wise for policy makers to segregate disruptive students, as these students generate the largest negative spillover effects on their peers, and they are the ones benefiting the most from segregated environments.

## 2.6 Conclusion

The present study sheds light on short-term and long-term returns to SpEd programs for students with special needs. Using recent methodological developments in computational text analysis and in causal machine learning, this study leverages psychologists' written reports to address the problem of confounding in the absence of experimental design. My study complements the literature by showing that SpEd should not be considered as a single intervention. Each program differs in its expected returns, and inclusive programs are quite effective at generating academic success and labor market integration. More specifically, I find that returns to SpEd programs in inclusive settings (counseling and individual therapies) are positive for academic performance, and that tutoring has no measurable effect. When compared to receiving no SpEd, all inclusive treatments have zero to positive returns.

Moreover, this study contributes to our understanding of inclusive and segregated measures for students with SEN. In general, inclusion pays off in terms of academic performance, labor market participation and earnings in comparison to semi-segregation. The results of this study show that inclusion works better for (almost) all students with SEN. However, semi-segregation is the least detrimental for students with SEN who exhibit disruptive tendencies, or for nonnative students with SEN. Semi-segregation works best in schools with an lower average SES status and more nonnative speakers. Results presented in this study, however, do not extend to the analysis of full segregation, as students placed in fully segregated settings have almost no overlapping characteristics with students in semi-segregated and inclusive settings. Moreover, higher attrition and selection into test participation for students placed in fully segregated school environments make the assessment of academic returns difficult.

The optimal policy allocations analyses offer further insights into inclusive and segregated SpEd programs for policy makers. With the help of policy trees, I propose placement recommendations to improve aggregate school performance and better labor market integration. These policies are outcome-improving for students with SEN, and cost-reducing. By implementing such optimal policies, a policy maker could significantly increase average school performance and labor market integration for students in semi-segregation by reallocating them to mainstream classrooms. However, this reallocation would incur costs on their peers in mainstream classrooms. In light of the results from the heterogeneity analysis and from the reallocation exercise, it might seem wise for policy makers to segregate disruptive students, as these students generate the largest negative spillover effects on their peers, and they are the ones benefiting the most from segregated environments. However, it is ultimately up to the policy maker and to society at large to decide whether a small decrease in school performance is a cost that should be endured by mainstream students in order to significantly increase the performance of reallocated students.

This paper invites further research on two fronts. On the one hand, the role of teachers and SpEd teachers in decisions to implement inclusion or segregation must be further investigated. Teachers' input and teachers' productivity in inclusive and segregated settings are important factors of the success of inclusive policies. However, little is known about their value-added for students with SEN. On the other hand, inclusion has benefits that extend beyond measurable academic performance and labor market integration. For instance, inclusion is likely to affect non-cognitive skills such as altruism, self-esteem, and self-image. These aspects need to be investigated to deliver a more complete picture of inclusion. Finally, this paper is an invitation for further methodological research in optimal policy allocation and in the use of text as covariate.

# 3 | High-Ability Influencers?

# The Heterogeneous Effects of Gifted Classmates[*]

Simone Balestra (University of St. Gallen and CESifo)
Aurélien Sallin (University of St. Gallen)
Stefan C. Wolter (University of Bern, CESifo, IZA, SCCRE)

This chapter is forthcoming in the *Journal of Human Resources*.

*We study the causal impact of intellectually gifted students on their nongifted classmates' school achievement, enrollment in post-compulsory education, and occupational choices. Using student-level administrative and psychological data, we find a positive effect of exposure to gifted students on peers' school achievement in both math and language. This impact is heterogeneous: larger effects are observed among male students and high achievers; female students benefit primarily from female gifted students; effects are driven by gifted students not diagnosed with emotional or behavioral disorders. Exposure to gifted students increases the likelihood of choosing a selective academic track and occupations in STEM fields.*

**Keywords**: gifted students, peer quality, gender, math, peer effects
**JEL Classification**: I21, I24, I26, J24

[*] We are grateful to the editor and two anonymous referees for helpful comments. We thank Mirjam Bächli, Uschi Backes-Gellner, Eric Bettinger, Anne Brenøe, Jeff Denning, Beatrix Eugster, David Figlio, Guido Heineck, Immanuel Lampe, Edward Lazear, Helge Liebert, Joan Llull, Fanny Puljic, Juanna Schrøter Joensen, Ulrike Unterhofer, Martina Viarengo, Fabrizio Zilibotti, and Ulf Zölitz for their constructive comments. The paper benefited from valuable feedback at the following meetings: Bildungsökonomischer Ausschuss im Verein für Socialpolitik, European Association of Labour Economists, Ski & Labor Seminar, Swiss Society of Economics and Statistics, Verein für Socialpolitik, and Young Swiss Economists Meeting. For the data, we thank the School Psychological Service, the Ministry of Education of the canton of St. Gallen, and the Stellwerk test service provider. We also thank Walter Goetze (*bfb Büro für Bildungsfragen AG*) for providing us the exact values of the apprenticeship requirement level ratings. We acknowledge financing from the Swiss National Science Foundation (grant no. 176381).

## 3.1 Introduction

In a context where the inclusion of special-needs students in the main classroom (“mainstreaming”) is becoming the norm and where special education programs are increasingly being abandoned, evidence on the effects of inclusion on students’ well-being, achievement, and post-education opportunities is more needed than ever. One particular population, traditionally segregated into special education classes, needs to be thoroughly investigated in a mainstreaming context: *gifted students* – i.e., students with an intellectual ability significantly higher than average.[42] It is a priori unclear whether and where such students exert a positive influence on their classmates, where – feeling bored and not fitting in – they are perceived as disruptive elements, and where they have no discernible effect on their peers. The aim of the present paper is to resolve this question by examining if and how gifted students affect their non-gifted classmates’ achievement in secondary school and enrollment in post-compulsory education. Given the heterogeneous nature of peer effects in the classroom[43], our research emphasizes how the influence of gifted students differs for their male and female, high-achieving and low-achieving peers in math and non-math school subjects.

We analyze the impact of gifted students on their classroom peers in the context of the Swiss education system, an inclusive academic setting which offers ideal conditions for the identification of spillover effects. One such feature is that no gifted students are segregated into gifted programs and that they are all included in regular schools, even though they may receive additional services or activities outside of the classroom. A second feature is that the status of gifted students is assessed and determined by the school psychological service, an independent and centralized institution that provides students and their families with diagnosis and counseling for school-related issues. This practice ensures that professional psychologists (and not parents, teachers, or school administrators) diagnose students as gifted. This in turn allows us to differentiate gifted students from simply high-achieving students and to break the myth that giftedness is the same as high achievement. In fact, nearly 50%

[42] We understand gifted children or students as “Children, students or youth who give evidence of high performance capability in areas such as intellectual, creative, artistic, or leadership capacity, or in specific academic fields, and who require services or activities not ordinarily provided by the school in order to fully develop such capabilities." (US Federal government statutory definition of gifted students, P.L. 103-382, Title XIV, p. 388).

[43] As pointed out by Booij, Leuven, and Oosterbeek (2017); Hoxby (2000); Lavy, Silva, and Weinhardt (2012); Lavy, Paserman, and Schlosser (2012); Burke and Sass (2013); Black, Devereaux, and Salvanes (2013).

of gifted students in our sample are not in the top five students of their class. We use student-level administrative data on achievement combined with detailed psychological examination records, uniquely linking students' school performance in a compulsory standardized test and administrative records from the school psychological service for ten consecutive cohorts of eight graders. To investigate career trajectories, we merge our data with administrative records containing detailed information on students' post-compulsory education choices.

For identification, we rely on the variation in classroom composition arising from within-school assignment of gifted students to classes when students transition from primary school (grades one through six) to secondary school (grades seven through nine). When transitioning from primary school to secondary school, students are assigned to new classes in their new school and remain in the same class for the rest of their mandatory education. For equity reasons and to avoid stigma, information on students' psychological profiles is usually not shared between primary and secondary schools. This practice implies that gifted students can be neither identified nor assigned to specific classrooms or teachers as students enter secondary school. We demonstrate, using several tests, that the observed within-school, between-class variation in the proportion of gifted classmates is consistent with variation generated from a random process. We also find no systematic assignment of gifted students to a specific class or teacher. Causal conclusions are further motivated by the fact that parents do not have a free choice of school in Switzerland: nearly all pupils attend public schools, and the geographical distribution of the population in the individual communities shows no regularities that would correlate with the distribution of talented pupils. Finally, the low prevalence of intellectual giftedness (approximately 2% of the population) is useful for the estimation of peer effects, allowing us to conduct the analysis exclusively on non-gifted students without losing a significant portion of the sample. By excluding gifted students from the analysis, we explicitly distinguish between the subjects of a peer effects investigation (regular students) and the peers who potentially provide the mechanism for causal effects on these subjects (gifted students).

The results are the following: we document a positive effect of exposure to students identified as gifted in all school subjects. Our results indicate that exposure to gifted students on average raises achievement of the other students by 8.7% of a standard deviation in math and 7.8% in language. Exposure to gifted students is daily classroom exposure over two school years (grade 7 and grade 8). When looking at who benefits the most from the presence of gifted students in the classroom, we observe the strongest effect for male students and for high achievers. In addition, we uncover a clear effect heterogeneity across gender and school subject. In math, male students profit significantly more than female students from gifted classmates, amplifying the gender gap in math achievement by 16%. In contrast, we find no significant gender difference in the spillover effect for language. This gender-subject heterogeneity is quite striking because classrooms (and thus peer composition) remain the same for all subjects. Moreover, we detect no other significant effect heterogeneity for characteristics like student's age, their native speaker status, class size, or teacher's gender.

Which gifted students generate the positive externalities? We provide compelling evidence that both the gender and the behavior of the gifted students matter, and that they matter even more to female students. We show that male students benefit from the presence of gifted peers in all subjects regardless of the gender of the gifted, whereas female students benefit almost exclusively from gifted female students. This pattern is more apparent in math and suggests that exposure to high-ability female peers may provide female students with a role model in quantitative fields, alleviating the negative effects of gender stereotypes. Not every gifted student is, however, a good peer. By distinguishing between gifted students who suffer from behavioral, emotional, or social problems from the other gifted students, we are able to isolate gifted students who do not exhibit disruptive behavior in the classroom. We find that female students are negatively affected by the presence of classmates who are gifted but disruptive. The evidence suggests that well-behaved gifted students improve their classmates' performance through both ability spillovers and reduced classroom disruption.

In terms of human capital investment, we further analyze students' career trajectories after compulsory education. In Switzerland, students after compulsory education must choose between vocational training or pre-university education (commonly known as "academic track") that provides them with the required skills to study at the tertiary level. By looking at whether students choose an academic track or a vocational track, we find that being exposed to gifted classmates in secondary school significantly increases the likelihood of choosing the academic track. This effect is entirely driven by male students who enter the academic track instead of the vocational track, which reflects the main findings and may offer an additional explanation to the persistent under-representation of women in math-intensive careers (e.g., in STEM fields). We investigate this hypothesis by classifying each vocational occupation according to its STEM content. We find that exposure to gifted students increases the likelihood of choosing an occupation in STEM fields among students entering the vocational track, an effect observed only among men.

The present paper contributes to and brings together three strands of literature in economics. First, we contribute to the under-investigated field of research on gifted students. Rather than looking at how gifted students perform when they are segregated into talented programs (e.g., Bui, Craig, and Imberman, 2014; Booij, Haan, and Plug, 2016), we bring evidence on the situation of gifted students in an inclusive education system, and we propose a new approach to identify high-ability students in general. The literature so far has used previous achievement (e.g., Booij, Leuven, and Oosterbeek, 2017), individual fixed effects (e.g., Burke and Sass, 2013), socioeconomic background (e.g., Black, Devereaux, and Salvanes, 2013), and parents' education (e.g., Cools, Fernández, and Patacchini, 2019) to determine student ability. Instead, we use formal assessments by external specialists (school psychologists) to identify gifted students. These external assessments are reliable assessment of students' cognitive abilities, extend beyond pure school performance and are less prone to biases arising from parents, teachers, or developmental factors. Given that being diagnosed as a "gifted" student does not automatically determine eligibility to targeted academic programs in Switzerland, our hybrid measure based on specialists' assessment and IQ tests is less likely to be manipulated.

Second, we contribute to the extensive literature on peer effects in education. This literature has provided quantified evidence that educational success cannot be explained only by students' own characteristics, parental background, and school environment, but that peers and the interactions between peers matter. One novel feature of our study is that, among the many peer dynamics occurring in the classroom documented so far, the heterogeneous influence of the population of gifted students has never been investigated. In addition, we are able to observe the classroom environment, where teaching occurs and students presumably affect directly their peers' learning. Although many scholars agree that classroom interactions play an important role in determining students' academic achievement and in shaping students' educational choices, most studies define peers at the school or cohort level. This definition of peer group may miss important interactions within classroom groups, because the estimation of spillover effects differs depending on the accuracy with which one identifies the set of relevant peers (Carrell, Sacerdote, and West, 2013; Carrell, Fullerton, and West, 2009). Finally, by presenting evidence on both school performance and career trajectory, we complement an emerging literature examining how peer characteristics during adolescence influence later career choices (Mouganie and Wang, 2020; Card and Payne, 2017; Zölitz and Feld, 2021; Anelli and Peri, 2017; Carrell, Hoekstra, and Kuka, 2018; Black, Devereaux, and Salvanes, 2013).

Third, we contribute to a growing strand of literature that aims to understand the roots of the persistent gender gap in math (for a recent review, see Buckles, 2019). Although the gender gap in education enrollment and labor market participation has dramatically narrowed over the past 50 years, the gender gap in math achievement still persists in most developed countries (Ellison and Swanson, 2010).[44] The reasons for this persistence are still not totally understood: recent research shows that the gender gap in math achievement does not exist upon entry to school, supporting the idea that nurture (e.g., gender stereotypes, culture) rather than nature (e.g., innate biological differences between sexes) determines gender differences in achievement (Pope and Sydnor, 2010; Hyde and Mertz, 2009; Nosek et al., 2009). However, the gap appears to be large and significant in the middle school years and beyond (Fryer and Levitt, 2010), and is in turn mirrored in the education and career choices of young women (Card and Payne, 2017; Carrell, Page, and West, 2010; Brenøe and Zölitz, 2020). It is therefore crucial to understand what are the factors in the school environment that originate and widen the gender gap in math achievement, especially for students at the age of choosing their first important career direction. This study offers both new evidence on the formation of the gender gap in math and the likely mechanisms behind such gap.

## 3.2 Background and Data

### 3.2.1 Institutional Background

The education system in Switzerland has a federal structure and gives the cantons – similar to the states in the U.S., the countries in Germany, or the provinces in Canada – great freedom in educational policy decision-making. In contrast to the other three federal states, however, the degree of coordination between the cantons is relatively high and, depending on the language region (German, French, or Italian), the cantons now apply the same common curriculum in all subjects. In the Intercantonal Agreement on the Harmonization of Compulsory Education, which the majority of

[44] Data from the 2015 PISA study reveal that Switzerland has one of the highest gender gaps in math performance (2.3%), alongside with other countries exhibiting above-average gender gaps like the United States (1.9%), the United Kingdom (2.2%), and Germany (3.1%).

the cantons – including the one we consider in the present paper – have joined, equal school structures were established. These include a two-year entry level (kindergarten) and nine years of compulsory schooling, of which the first six years are allocated to the primary level and the last three to the lower secondary level. Pupils change schools and classes when moving from primary to lower secondary education.

Within each canton, schools are organized at the municipality level, and children are assigned to schools on the pure basis of their location of residence. This strict assignment procedure is thoroughly implemented, such that parents have no say about their child's school other than moving permanently to a different municipality or enrolling their children in a private school. Despite this rule, private schooling remains very rare in Switzerland. As the 2018 Education Report by the Swiss Coordination Center for Research in Education shows, more than 95% of children in 2016 attend public-funded schools in their municipality of residence.

The present analysis focuses on all students enrolled in the secondary schools of the Canton of St. Gallen (around 500,000 inhabitants). In this state, children are required to undergo eleven years of compulsory education, divided into kindergarten (two years), primary school (six years), and secondary school (three years). In most cases, secondary schooling takes place in larger schools administered by associations of municipalities (districts). Tracking occurs at the secondary school level and is based on students' academic performance in primary school as well as their teacher's recommendation. Students are either sent to a high-track secondary school (*Sekundarschule*) or to a lower, more practice-oriented secondary school track (*Realschule*). Once allocated to one of the two secondary school tracks, students are assigned to classes within each school-track. The administrative staff of each school has no prior knowledge about the students other than administrative data on gender, place of residence, primary school attended, and nationality. For equity reasons, information on students' disabilities, special needs, or high ability status is usually not shared between primary schools and secondary schools. This practice potentially creates a situation where class composition is quasi-random with respect to students' psychological profiles, a situation that we evaluate in the empirical strategy section. Once assigned to a class, students share the same peers for all the lectures and subjects, and classes remain unchanged for the three years of secondary school.[45]

[45] Grade repetition is not common in St. Gallen, as only about 1.5% of students in a cohort ever repeated a grade. Within-state and out-of-state mobility in St. Gallen are also low at about 2-3% (data are from the Swiss Federal Statistical Office for the years 2008-2017).

At the end of their eighth year of compulsory schooling (second year of secondary school), all students are subjected to a mandatory standardized test (the so-called “Stellwerk 8" test). This computer-based adaptive test automatically adapts the difficulty of questions to the ability and knowledge revealed by the student in the previous questions (in the same fashion as the GMAT or GRE test). It tests core knowledge of mathematics, language (German), and, depending on the track, foreign languages (usually English) as well as natural sciences (including biology, chemistry, or physics). The correction of the test is computer-based, which eliminates concerns of teachers' bias or stereotyping in the results. The results are important both for students, who will use the test scores when choosing their post-compulsory education, and teachers, whose relative performance can be reflected by the rate of success of their students. Although Stellwerk 8 is not needed to obtain the compulsory school diploma, students receive a certificate with their results and usually submit it to potential employers when applying for VET positions.[46]

The Canton of St. Gallen bears the responsibility for the inclusion and education of children with high abilities and must guarantee the fulfillment of their educational needs. In this regard, emphasis is put on inclusion of students identified as gifted in regular classrooms (mainstreaming). Requests to send gifted children to special schools are accepted only under strict conditions: the child must have already skipped a grade, justify why the classroom environment is not adequate, and undergo a psychological evaluation. In all other cases (the vast majority), special activities, additional support, and enrichment programs are offered outside of class, depending on the school and upon request by parents and teachers. Acceleration (skipping a grade) or school start at a younger age are also possible in some rare cases.

The task of identifying and providing psychological support to children with high abilities is carried by the School Psychological Service (SPS), a centralized and independent office. It provides diagnoses of learning disabilities, behavioral difficulties and developmental deficiencies, assigns therapies and treatments to children, and offers counseling to children, parents, and teachers. For most students (about nine out of ten), services of the SPS are requested directly by the teacher, but some requests are also filed by the parents or the child's doctor. The referring party needs to justify its

[46] The Stellwerk test is used in many German-speaking cantons and Stellwerk scores are also used to describe occupation profiles in the Swiss German labor market. For example, one of the most common occupations in Switzerland – commercial employee – requires a Math score between 425 and 525, a German score between 550 and 650, and an English score of between 550 and 650.

request by pointing out the reason for the child's registration with the SPS. The reasons most commonly brought up are learning disabilities, social or emotional problems, difficulties with the family and the parents, or challenging relationships with the teacher. After a request has been made, children and their parents are directly contacted by a caseworker from the SPS for an assessment of the situation and a health diagnosis. As part of the diagnosis, an intelligence test is often administered to children.

### 3.2.2 Data Sources

We use information on classroom composition, characteristics of students, and individual academic achievement from the Stellwerk test taken in eighth grade by the entire population of students in the Canton of St. Gallen.[47] To this, we add information from the administrative records of the SPS on individual psychological profile, giftedness status, and learning disabilities of each child. After merging these two sources, we observe the academic achievement, psychological profile, and peer group composition for each student enrolled in grade eight in the Canton of St. Gallen for ten consecutive school cohorts (2008 to 2017).

More precisely, the test score data allow us to observe the following for each cohort: composition of secondary school classes (with the school, the track, and the classroom as well as the teacher ID), basic characteristics of each student (birth date, gender, and whether the student is a native German speaker) and student academic achievement on the Stellwerk test (scores for all examined materials). In this analysis, we focus primarily on the scores in math, language (German), and a composite of the two, which are compulsory subjects for all students in all tracks, and standardize them with mean zero and standard deviation one. As we mentioned in the previous section, the classroom composition we observe in the Stellwerk data is the classroom composition that remains fixed over the whole three years of secondary school in all subjects.

Information on the gifted status of students is given by the administrative records of the SPS. In these records, we find information on each child who has had contact with the SPS at any point in his or her school years. They contain the reason of registration, the therapies assigned to the child, the number of visits to the SPS, the date of each visit, and all the notes left by the caseworker about the child. These notes give a very detailed source of information about the child's situation, the topics dis-

[47] This section follows Balestra, Eugster, and Liebert (forthcoming) closely, who use an extended version of the same data set.

cussed during each interview, and an overall idea about the diagnosis. Important for our study, we observe the IQ score for many of the children registered at the SPS. Most of the requests to the SPS are made when the child is between six and nine years old, and first contacts with the SPS in primary school often coincide with the time when children start receiving school grades (second semester of second grade).

Four restrictions are imposed on the data, reported in detail in Appendix Table A.1. First, we restrict our data set to students enrolled in the higher track (*Sekundarschule*, 62% of the original sample) and discard those in the lower track (*Realschule*). The reason for this is that the vast majority of gifted children (93%) pursue their education in the Sekundarschule. One advantage of focusing on the higher track only is that the sample is very homogeneous with respect to ability. Second, we focus only on students who were actually required to take the Stellwerk test. This leaves out students from special education institutions, for which we do not observe complete classes. Third, we exclude segregated classes that are composed only of students with special needs. Finally, we remove classes and students with missing or implausible values (e.g., test scores exceeding the possible range, classes which are too small or large, or negative age at test). We are left with a final sample of 31,625 students in 1,592 classes from 80 schools.

### 3.2.3 Definition of the Key Variables

While remaining aware that intellectual giftedness is a multi-faceted concept whose definition has never been generally agreed upon (Sternberg, Jarvin, and Grigorenko, 2010), we understand intellectual giftedness as an intellectual ability significantly higher than average. Intellectual giftedness is believed to persist as a trait into adult life, with various consequences studied in longitudinal studies of giftedness over the last century (Gottfried et al., 1994). Albeit no generally agreed definition of giftedness for either children or adults has been reached, most school placement decisions and longitudinal studies over the course of individual lives have followed people with IQs in the top two percent of the population (Newman, 2008) – that is, IQ scores above 130 (two standard deviations above the mean). However, there is substantial variation in the threshold used across theories of intelligence, intelligence scales, and individual psychologists.[48]

[48] For instance, Silverman (2018) uses the threshold of 120 to identify "mildly gifted," 130 for "moderately gifted," 145 for "highly gifted," 160 for "exceptionally gifted" and 175 for "profoundly gifted."

IQ scores are known to mildly predict academic achievement (Neisser et al., 1996; Deary et al., 2007), since school success is also strongly determined by dedication, motivation, and parental background and investment. Criticism within the psychological community has raised doubts on the validity of the IQ score (see discussion in Sternberg, Jarvin, and Grigorenko, 2010): the IQ score, which maps intelligence unidimensionally, might miss other cognitive dimensions relevant to intelligence, such as emotional intelligence (Mayer et al., 2001; Zeidner et al., 2005), creativity (although much debated, see Make and Plucker, 2018), or domain-specific abilities. Nonetheless, the advantages of measuring cognitive ability with a uniform, normalized IQ scale and of tying the definition of "giftedness" to a particular threshold score on the IQ scale are manifold, such as psychometric advantages (easy quantification of intelligence, reliability, internal and test-retest consistency), transparency (the concept of IQ is widely known), predictive accuracy for intelligence in general (e.g., Der, Batty, and Deary, 2009), and external validity (measures of IQ exist for all ages, and have been normed across cultures and countries, e.g., Lynn and Meisenberg, 2010; Lynn and Vanhanen, 2012). As a consequence, in the US for instance, individual IQ testing is becoming less commonly used for identification of the gifted and a more holistic identification procedure is preferred.[49]

Our data allow us to mitigate the potential limits of IQ as a unique and reliable measure of giftedness: since the SPS records not only report the children's IQ scores, but also qualitative assessments of cognitive ability (as obtained from the diagnoses and comments of the caseworker), we are able to enhance the IQ scores with qualitative assessments. Qualitative assessments reliably complement quantitative scores, take into account other dimensions of intelligence not assessed by the IQ test, and allow for discarding false positives (Silverman, 2018).[50] They also allow us to differentiate between high achievers and gifted students, as the two do not necessarily overlap. In comparison to the previous literature, we can integrate a richer notion of high intelligence in our analysis.

We then proceed in the following three steps to construct our indicator of giftedness. First, we select one IQ score per child. For many children, we observe different

Moreover, the scale must be adapted for students with a native language different than the one of their main environment. Some researchers also use an IQ threshold of 116 for non-native speakers (Card and Giuliano, 2016).

[49] See the report from the National Association for Gifted Children, 2015, `www.nagc.org`, and Peters and Matthews (2016).

[50] We are aware of the existence of bias in qualitative assessments, such as documented in McDermott,

scores, either taken at different points in time, and/or estimated with different intelligence tests. For each child with more than one score, we take the highest score reached by the child.[51] With respect to the chosen intelligence test, we know from the SPS that each child is given the test that suits his or her situation the best.[52] We classify the child as gifted if his or her IQ score is equal to or above 130. We conduct our main analysis with a threshold of 130, and we show that our results are robust when applying more or less restrictive definitions of giftedness. In a second step, we code the written diagnosis of the caseworker – a trained psychologist – and assign the gifted status to the children who are diagnosed as such by the caseworker. In a final step, we remove false positives, i.e., children with a high IQ score but whose assessment does not diagnose high ability. For example, we discard cases in which the child reaches a high IQ score, but the caseworker writes that the child had learned how to perform well on the test from his or her siblings. It is important to mention that we focus exclusively on gifted students who are identified prior to secondary school entry. By doing so, we make sure that gifted status does not depend on class composition in secondary school.

In sum, 20.5% of all our observations have been referred to the SPS and 12.8% of all students have been assessed for IQ with a formal IQ test. Of the 578 students classified as gifted, 145 students were identified as gifted with both measures, IQ test and diagnosis, whereas 329 students were identified only with the diagnosis and 104 exclusively with the IQ test. Finally, there are 82 students who were referred to the SPS for an assessment of giftedness but were not diagnosed as gifted. Our measure of giftedness has never been used in the literature on ability peer effects and we argue that using a metric based on experts' diagnoses is less prone to misreporting (e.g., self-assessment), context-specific factors (e.g., school or class composition), and external influences (e.g., parents or teachers).

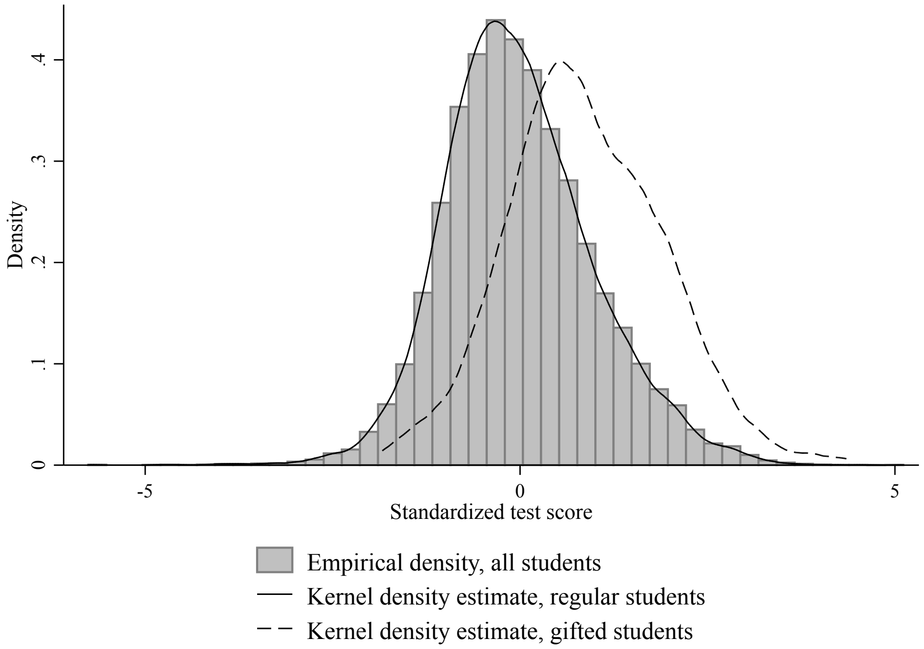
### 3.2.4 Summary Statistics

Table 3.1 reports the summary statistics for our final sample. The typical eighth grade class consists of 20 students and there are 0.36 gifted classmates per class. De-

Watkins, and Rhoad (2014). Unfortunately, we do not observe the caseworker ID so we cannot take bias into account in our analysis.

[51] After discussion with the head of the SPS, we decided to take the highest score since many children need more than one attempt to be fully concentrated during the test. Using either the first IQ administered or the average of all IQ tests performed has no substantial impact on the results.

[52] The available intelligence tests are Snijders-Oomen nonverbal intelligence tests (SON), Kaufman Assessment Battery for Children I and II (K-ABC and K-ABCII), the Wechsler Intelligence Scale for Chil-



*Notes:* Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Figure 3.1: Distribution of test scores

spite this low prevalence, 27% of all students are exposed to at least one gifted classmate in eighth grade. Every second student is a female student, one in ten students is non-native German speaker, and the average age at which students take the Stellwerk test is approximately 15.

The subsample of 578 gifted students is of particular interest, as shown in Table 3.1, Panel D. Around 1.9% of the sample is identified as gifted, which is slightly below the expected percentage of individuals with $IQ \geq 130$ under the normal curve ($\approx$ 2.1%). This discrepancy is likely explained by the fact that some gifted individuals undergo primary school undetected, thus never entering in contact with the SPS before reaching secondary school (for identification purposes, we only focus on gifted students identified during primary school). Consequently, as we only observe *identified* gifted students, our measurements of being exposed to gifted students on other students will underestimate the true effect (attenuation bias). In this sense, our findings can be interpreted only as the effect of being exposed to students identified as gifted.

dren (WISC-Hawik), Raven's Progressive Matrices (Raven), Kramer-Test, and Culture Fair Intelligence Test (CFT).

| | (1)<br>Sample mean | (2)<br>Standard deviation |
|---|---|---|
| **A. Outcome:** | | |
| Test score: Mathematics | 0.000 | 1.000 |
| Test score: Language (German) | 0.000 | 1.000 |
| Test score: Composite | 0.000 | 1.000 |
| **B. Exposure to gifted children:** | | |
| Gifted student | 0.018 | 0.134 |
| Gifted classmates (proportion) | 0.018 | 0.035 |
| Gifted classmates (number) | 0.351 | 0.673 |
| Exposure to gifted classmate | 0.264 | 0.441 |
| **C. Classroom characteristics:** | | |
| Female | 0.524 | 0.499 |
| Native German speaker | 0.912 | 0.284 |
| Age at test | 14.81 | 0.699 |
| Class size | 20.52 | 3.378 |
| Male teacher | 0.649 | 0.477 |
| **D. Characteristics of students identified as gifted (**$N = 578$**)** | | |
| Female | 0.296 | 0.457 |
| With emotional or social problems | 0.123 | 0.329 |
| Referred to SPS by teachers | 0.856 | 0.351 |
| Referred to SPS by parents | 0.126 | 0.332 |
| Rank in the class (composite test) | 6.242 | 5.238 |

*Notes:* Descriptive statistics for the main estimation sample, based on 31,765 students in 1,597 classes from 80 schools. In regressions analyses we exclude gifted students from the estimation sample, which reduces the number of students to 31,187. Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Table 3.1: Descriptive statistics

Approximately 30% of the identified gifted students in our sample are female, which is similar in proportion to other European countries.[53] The under-representation of gifted female students, documented in many different contexts (Petersen, 2013), might be explained by teachers' gender bias in referrals for giftedness assessment (which has been documented by Bianco et al., 2011) or by the fact that gifted female students are less likely to be identified by means of IQ testing and standardized tests (Petersen, 2013). It can also be the consequence of a higher prevalence of disruptive behaviors (and thus a higher prevalence of references to the SPS) among male students in general, which makes female students pass unnoticed by the teachers and not assessed by the psychologist: in our data, while 25% of male students have been referred to the SPS in general, only 16.5% of female students were referred.

Appendix Figure A.1 shows the distribution of class size and gifted classmates in absolute and relative terms. The figure indicates that while most classes have no gifted student, nearly all classes exposed to gifted students contain exactly one gifted student. Figure 3.1 exhibits the distribution of test scores by student type, showing that gifted students perform on average better than regular students – almost a standard deviation better. However, not all gifted students are high achievers and, at the same time, not every high achiever is classified as a gifted student. This finding is confirmed in Appendix Figure A.2, which shows the distribution of gifted students' classroom ranks on the Stellwerk test and reveals every second gifted student performs in the top five of his or her classroom, while the rest might perform even in the lowest ranks. This is likely because our definition of gifted transcends the test score dimension, featuring a measure of intellectual ability based on psychological examinations rather than previous achievement.

## 3.3 Empirical Strategy

The aim of this paper is to evaluate the impact of exposure to identified gifted peers on student achievement. Empirically, we estimate the following linear model:

$$Y_{icst} = \alpha + \beta \, \text{Exposure}_{cst} + \gamma \, X_{icst} + \delta \, \overline{X}_{(-i)cst} + \varepsilon_{icst} \tag{3.1}$$

[53] Data are from the EASIE 2014 Dataset Cross-Country Report, retrieved from `https://www.european-agency.org/data`.

where $Y_{icst}$ is the outcome of interest, such as the math test score of student $i$ in classroom $c$, school $s$, and year $t$. $X_{icst}$ is a vector of individual characteristics that include age at test, an indicator for gender, and an indicator for native German speaker. $\overline{X}_{(-i)cst}$ is a vector of average characteristics of $i$'s class (class size, proportion of female students, proportion of native German speakers, and mean age). The variable of interest is Exposure$_{cst}$, a measure of exposure to *identified* gifted students in a given class. We parametrize such measure for each student as an indicator being exposed to at least one gifted classmate in grade eight, but the results are consistent to alternative specifications (e.g., the proportion of gifted students per class). The peer spillover parameter is $\beta$, which represents the impact of being exposed to a gifted student on $i$'s outcome. The error term $\varepsilon_{icst}$ is assumed to consist of two components: a school-by-year fixed effect and an idiosyncratic error term (i.e., $\varepsilon_{icst} = \mu_{st} + e_{icst}$). Finally, standard errors are clustered at the classroom level throughout the paper.

The estimation of the interest parameter $\beta$ suffers from three main identification problems. First, Manski (1993)'s well-known reflection problem states that all behaviors in a peer group are affected by the behaviors of the other members of the group. Namely, a student simultaneously influences the outcome of the group and the group influences the outcome of the student. We tackle this problem in two ways. First, all variables in Equation (3.1) are determined before secondary school, including the status of gifted student. This strategy ensures that neither the gifted status nor other individual characteristics are influenced by contemporary class composition.[54] Second, we exclude gifted students from all regressions in order to separate the subjects of our investigation (regular students) and the peers who potentially provide the mechanism for causal effects on these subjects (gifted students). As discussed by Angrist (2014), this distinction eliminates mechanical links between own and peer characteristics, making it easier to isolate variation in peer characteristics that is independent of subjects' own characteristics.

The second main identification problem stems from common unobserved shocks at the group level. These shocks at the class and school levels could tamper with the

[54] Our measure of exposure could be weakened if students were regularly changing classes in secondary school. However, student mobility between schools is rare. Data from the official education statistics from the Swiss Federal Statistical Office (*Statistik der Lernenden* in German) indicate that, in the state of St. Gallen, approximately 40 students per year change school between grade seven and eight. This figure corresponds to a prevalence of 0.77%, and prevalence never exceeds 1% in the years considered (2011-2016). Note that school mobility is primarily due to students moving to another municipality, which accounts for 70% of total school mobility.

identification of peer effects of gifted students on their classmates. For instance, the outbreak of an epidemic or the introduction of new pedagogical methodologies for a lesson could impact the overall academic performance of a classroom or a school, which would confound the peer effects estimation if correlated with the proportion of gifted peers. To resolve this issue, we introduce a series of fixed effects that control for unobserved heterogeneity at multiple levels (namely the school-by-year level).

The third identification problem is endogenous peer selection and is the most challenging to tackle. If individuals are systematically assigned to groups according to a specific characteristic, the researcher cannot determine whether a difference in outcome is a causal peer effect or simply an artifact of due to group assignment. We take care of this problem by ensuring that gifted students are quasi-randomly assigned to classes at the secondary school level (i.e., identification between classes within the same school-year). As we already mentioned previously, the transfer from primary to secondary school is regulated in such a way that students are assigned to schools based on their place of residence. In each school, students from different places of residence (and consequently different primary schools) are mixed. Importantly, students' psychological profiles are unknown to secondary school administrators such that equity among students is guaranteed and stigma when transitioning between schools is avoided. We exploit this policy rule for identification and formally test the validity of the strategy with three types of balancing tests.

We begin by testing whether individual and group characteristics predict exposure to gifted peers. The aim of this test is to detect potential selection into classrooms. Table 3.2 shows the results, where each regression includes school-year fixed effects. Panel A focuses on individual level characteristics (gender, native German speaker, and age), while panel B focuses on classroom characteristics (proportion of female classmates, proportion of classmates who native speakers, classmates' mean age at test, and class size). None of the coefficients in Table 3.2 are statistically significant and, in addition, the size of the coefficients is very small. We also test for joint significance of the individual characteristics (panel A, column 4) and group characteristics (panel B, column 5). We cannot reject the null hypothesis that the coefficients on gender, native speaker, and age are jointly zero, with a p-value of 0.550. Neither can we reject the null hypothesis that the coefficients on the class-level characteristics are jointly zero (p-value of 0.449).

| | (1)<br>Exposure to gifted peers | (2)<br>Exposure to gifted peers | (3)<br>Exposure to gifted peers | (4)<br>Exposure to gifted peers | (5)<br>Exposure to gifted peers |
|---|---|---|---|---|---|
| | A. Individual characteristics | | | | |
| Female | 0.004 | | | 0.004 | |
| | (0.003) | | | (0.003) | |
| Native speaker | | 0.002 | | 0.002 | |
| | | (0.009) | | (0.009) | |
| Age at test | | | -0.001 | -0.001 | |
| | | | (0.003) | (0.003) | |
| *Joint significance (p-value)* | | | | *0.550* | |
| | B. Classroom characteristics | | | | |
| Female classmates (%) | -0.119 | | | | -0.101 |
| | (0.172) | | | | (0.172) |
| Native speaker classmates (%) | | 0.115 | | | 0.134 |
| | | (0.150) | | | (0.152) |
| Classmates mean age at test | | | -0.042 | | -0.044 |
| | | | (0.037) | | (0.036) |
| Class size | | | | 0.010 | 0.010 |
| | | | | (0.008) | (0.008) |
| *Joint significance (p-value)* | | | | | *0.449* |
| School-by-year FE | Yes | Yes | Yes | Yes | Yes |
| Observations | 31,187 | 31,187 | 31,187 | 31,187 | 31,187 |

*Notes:* *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$. Standard errors, shown in parentheses, are clustered at the school-year level (level of randomization). Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Table 3.2: Balancing tests

Although we find no evidence for selection into classrooms according to observable characteristics, we might suspect selection into classrooms based on unobservables. We thus perform a direct test of our identification strategy that assignment of gifted students within school-years is as good as random. To do so, we follow closely the approach of Chetty et al. (2011) and regress the gifted indicator on both school-year fixed effects and class fixed effects. Then we test whether the class fixed effects are jointly significant. The coefficients on the class indicators estimate the difference in probability of being assigned to a given class relative to the reference classroom within the same school-year cell.[55] If these coefficients are not statistically different from zero, we conclude that the probability that a gifted student is assigned to a specific class within a given school-year is the same for all the classes of that school-year. The p-value of the F-test is 0.630, supporting the key identifying assumption that the observed variation in exposure to gifted students between classes of the same school-year is random.

We repeat the test presented in the last paragraph for teacher assignment. Because no teacher holds two classes in the same year, we separate the school and the year fixed effect, but the procedure remains the same. We regress the gifted indicator on school fixed effects, year fixed effects, and teacher fixed effects and test whether the teacher fixed effects are jointly significant. The resulting p-value is 0.285, suggesting that gifted students are not systematically assigned to certain teachers.

In sum, all the tests performed support the validity of our identification strategy. Further evidence on this is presented by Vardardottir (2015). Using PISA data from secondary schools in Switzerland, she shows that track-by-school fixed effects render peer group composition conditionally uncorrelated with a large set of students' characteristics, while track fixed effects and school fixed effects separately do not. The approach we follow in the present paper is even more conservative, as we exploit variation within school-(track)-year. Note also that while families can potentially choose their district of residence thereby influencing schooling options for their children, possible selection into schools does not confound the results. In any case, mobility in Switzerland is generally low: approximately 80% of people do not move within five years and moving for school choice alone is likely to be a rare occurrence. In addition, municipalities within the canton of St. Gallen are very homogeneous in terms of demographics, indicating that such strategic behavior is most likely limited. The between-

[55] Note that one classroom fixed effect within each school-year cell is always omitted to prevent multicollinearity between classroom and school-year fixed effects (see Chetty et al., 2011, pages 1609-1610).

municipality variation in the unemployment rate (coefficient of variation: 0.42), the share of rich (0.19) and poor taxpayers (0.69), and the share with secondary (0.19), higher secondary (0.07), and tertiary education (0.22) is small.[56] Finally, we find low geographical variation when examining the prevalence of gifted students across municipalities, as Appendix Figure A.3 shows.

Having established that variation across municipalities is low, one might be concerned that estimated effects would be biased if students exposed to gifted classmates in grade eight were more likely to have superior educational inputs in years prior to treatment. The best strategy to address this concern would be to control for prior achievement. This is, however, not feasible in our context, for two reasons. First, no standardized test prior to Stellwerk 8 exists; second, we have no information on primary school GPA in the data. Even if we had primary school grades or GPA, such measures would not be comparable across classes, because in Switzerland the responsibility for grading and assessing performance is left to the teacher. To resolve the issue of systematic differences in prior educational inputs by exposure to gifted students, we examine the correlation between exposure to gifted classmates and educational inputs at the municipality level. We use two measures of educational inputs previously used in the literature, namely per-student spending (e.g., Jackson, Johnson, and Persico, 2016) and socio-economic composition (e.g., Angrist and Lang, 2004). The data on spending per primary school students comes from the official accounts published by municipalities at the end of the fiscal year, whereas the information on the socio-economic composition of each municipality is provided by Competence Center for Statistics within the Department of Economic Affairs of the Canton of St. Gallen.[57] Appendix Figure A.4 presents the result for per-pupil spending, while Appendix Figure A.5 presents the results for socio-economic composition. Both figures clearly show a flat relation between either measure and exposure to gifted students and none of the two slopes is statistically significant. This evidence suggests – albeit indirectly – that students exposed to gifted classmates in grade eight did not have access to superior educational inputs.

One last concern related to our empirical strategy is selective attrition. If students exposed to gifted peers are more likely to be observed in the data, this might induce

[56] Data are provided by Eugster and Parchet (2019).

[57] According to the data, municipalities spend on average 10,160 Swiss Francs (approximately 11,140 USD) per primary school student. This figure is higher than the OECD average (8,733 USD) but comparable to the corresponding figure in the U.S. (11,319 USD), as the OECD documents (OECD, 2017). The information on municipalities' socio-economic composition consists of a "social index" calculated

bias in our estimated effects. To resolve this potential issue, we conduct an attrition analysis by regressing exposure to gifted classmates on the following five outcomes: missing test score in Math, missing test score in German, missing test score in both Math and German, missing information on post-compulsory education, missing information on occupation profile. Appendix Table A.2 presents the results and shows that exposure to gifted students does not significantly predict any of the outcomes analyzed. The table also reveals that attrition rates are generally low and even very low for test scores (around 0.4%). These results show that some attrition is present but that it is not related to the treatment, which alleviates any worry regarding selective attrition. Note that attrition in the post-compulsory data originates primarily from the fact that the last two cohorts in the data are simply too young to appear in such data.

## 3.4 Results

In this section, we present and discuss the results in three parts. First, we introduce the main results on the effect of exposure to gifted students on test scores and perform a comprehensive sensitivity analysis. Second, we proceed to investigate potential heterogeneous effects and mechanisms driving the main results. Third, we examine longer-term outcomes by estimating the effect of exposure to gifted peers on post-compulsory education trajectories.

### 3.4.1 Main Results

The main results are presented in Table 3.3. Specifications 1 to 3 show the effect of exposure to peers identified as gifted on the composite test score (math and language) for all students with different sets of added regression controls, whereas specifications 4 and 5 consider math and language test scores separately. As regression controls, we include student-level controls (column 2) and classroom-level controls (column 3). Crucial for our identification strategy, school-year fixed effects are added to all specifications. Standard errors are clustered at the classroom level.

The estimated coefficients of exposure to identified gifted peers consistently reveal a positive effect on students' own academic performance. The most conservative specification indicates that exposure to gifted students raises the achievement of the

and provided by the Competence Center for Statistics within the Department of Economic Affairs of the

| | (1)<br>Composite<br>test score | (2)<br>Composite<br>test score | (3)<br>Composite<br>test score | (4)<br>Math<br>test score | (5)<br>Language<br>test score |
|---|---|---|---|---|---|
| Exposure to gifted classmates | 0.093*** | 0.094*** | 0.095*** | 0.087*** | 0.078*** |
| | (0.023) | (0.022) | (0.022) | (0.022) | (0.020) |
| Female | | -0.199*** | -0.181*** | -0.353*** | 0.047*** |
| | | (0.012) | (0.015) | (0.014) | (0.014) |
| Native speaker | | 0.404*** | 0.406*** | 0.196*** | 0.512*** |
| | | (0.021) | (0.021) | (0.020) | (0.022) |
| Age at test | | -0.189*** | -0.189*** | -0.166*** | -0.161*** |
| | | (0.009) | (0.009) | (0.009) | (0.009) |
| Female classmates (%) | | | 0.470*** | 0.420*** | 0.392*** |
| | | | (0.170) | (0.162) | (0.148) |
| Native speaker classmates (%) | | | -0.149 | 0.025 | -0.288** |
| | | | (0.142) | (0.137) | (0.128) |
| Classmates mean age at test | | | -0.011 | -0.013 | -0.005 |
| | | | (0.037) | (0.039) | (0.031) |
| Class size | | | 0.003 | -0.002 | 0.008 |
| | | | (0.008) | (0.008) | (0.007) |
| School-by-year FE | Yes | Yes | Yes | Yes | Yes |
| Observations | 31,187 | 31,187 | 31,187 | 31,187 | 31,187 |

*Notes:* *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$. Standard errors, shown in parentheses, are clustered at the classroom level. Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Table 3.3: Spillovers from gifted classmates

other students by 9% of a standard deviation on average. All effects are statistically significant at the 1% level, and adding covariates does not substantially change the estimates. Note that because we exclude the gifted from the analysis, these results do not reflect the effect of giftedness on the gifted students themselves. Moreover, exposure measured in our results reflect daily exposure over two school years. In terms of individual characteristics, we find the well-documented results that female students score better in language but worse in math than male students and that non-native speakers perform on average worse than native speakers (especially in language). Class-level characteristic appear to be quite irrelevant, except for the proportion of female students in the class: having more female classmates increases student achievement in all subjects. This finding corroborates previous studies conducted in the United States (Whitmore, 2005; Hoxby, 2000) and Israel (Lavy and Schlosser, 2011a). We also find that a higher proportion of non-native speaker classmates reduces achievement in language (column 5).

In order to assess whether the exposure to gifted peers impacts the gender gap, we examine potential gender-related heterogeneous effects in Table 3.4. We estimate the effect of the exposure to gifted peers including an interaction between exposure and own gender. We reject the null hypothesis that the effect is similar for both female students and male students in math. The estimated coefficients show that not only female students perform on average less well than male students on the math test, but also that the presence of gifted peers in the classroom exacerbates this difference. In math, the positive impact of the exposure to a gifted peer almost completely disappears for female students. However, female students do benefit from gifted peers as much as male students when it comes to performance in language.

Quantifying our results in terms of the gender gap, we find that female students exposed to gifted students are better off than female students without such exposure. However, compared to male students, the results suggest that the exposure to gifted students increases the gender gap in math performance by 16% (figures from Table 3.4, column 2). This increase is due to a disproportional increase in male students' performance when exposed to gifted classmates. However, the interpretation of the result is subject to the following caveat. Since our descriptive evidence indicates that

Canton of St. Gallen. The social index is based on the following four indicators: ratio of foreigners with citizenship of non-German-speaking countries in the population group of 5-14-year-olds, share of unemployed in the 15-64-year-old permanent resident population, ratio of 5-14-year-olds dependent on social assistance to the 5-14-year-old population, quota of low-income households with 0-13-year-old children.

| | (1)<br>Composite<br>test score | (2)<br>Math<br>test score | (3)<br>Language<br>test score |
|---|---|---|---|
| Exposure to gifted classmates | 0.116*** | 0.115*** | 0.085*** |
| | (0.026) | (0.026) | (0.024) |
| Female | -0.170*** | -0.338*** | 0.050*** |
| | (0.017) | (0.016) | (0.016) |
| Exposure * Female | -0.039 | -0.055** | -0.013 |
| | (0.027) | (0.027) | (0.026) |
| Individual characteristics | Yes | Yes | Yes |
| Classroom characteristics | Yes | Yes | Yes |
| School-by-year FE | Yes | Yes | Yes |
| Observations | 31,187 | 31,187 | 31,187 |

*Notes:* *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$. Standard errors, shown in parentheses, are clustered at the classroom level. Individual characteristics include gender, native German speaker, and age at test. Classroom characteristics include class size, share of females, share of native German speakers, and average age at test. Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Table 3.4: Spillovers from gifted classmates by subject and gender

gifted female students are more likely to remain undetected than male gifted students, it is even more important to understand the effect we measure as the effect of *identified* gifted students.

The results are not sensitive to alternative specifications, treatment or outcome definitions, or identification strategies. More specifically, we conduct four sets of robustness checks. First, to make sure that the results are not driven by our definition of exposure to identified gifted students, we conduct the same analysis while defining exposure as the proportion of gifted classmates. Table A.3 in the Appendix shows that our main results hold, showing that adding one gifted student to a class of 20 would increase achievement by approximately 5% of a standard deviation. We also explore potential nonlinearities in the share of gifted students per class by adding quadratic and cubic transformations of the share of gifted students per class. We do not find any nonlinearities in the relationship between average test scores and the share of gifted peers.

Second, we check that our results are not sensitive to a specific definition of giftedness. As mentioned previously, the cutoff in IQ score that determines whether a student is classified as gifted is debated in the literature. For this reason, we conduct the analysis by considering other IQ thresholds. Figure A.6 in the Appendix displays the results when IQs of 135, 130, 125, 120, 115, and 110 are used as thresholds for classifying a child as gifted. In addition, we follow Card and Giuliano (2016) and use the threshold of 116 for non-native speakers (130 for native speakers). We find that all the alternative thresholds are within the 95%-confidence interval of the main estimate (IQ threshold of 130).[58] The results are also robust to using alternative definitions of giftedness, namely gifted students identified only by means of IQ testing (quantitative assessment) and gifted students identified only by means of qualitative assessment.

Third, we check whether the same patterns occur for academic performance in other subjects. Appendix Figure A.7 presents estimates of the main specification for performance in natural sciences (biology, chemistry, and physics) and foreign language (English) as outcomes. These findings reinforce our conclusion that a gender gap in performance for STEM-related fields exists: similar to math, female students perform relatively less well than male students in natural sciences and do not benefit as much from gifted peers for science-related subjects. In contrast, and similar to the pattern we found for language, female students do perform relatively better than male students in foreign languages. The evidence thus suggests that the pattern we uncover in the main analysis for math and (first) language also extends to other STEM and non-STEM subjects.

Fourth, teachers are often mentioned as crucial determinants of students' performance and preferences for particular subjects.[59] To test whether the gender heterogeneity documented in the main specification arises from teachers' own characteristics (either observed or unobserved), we repeat the main analysis by adding teacher fixed effects. By doing so, we change our strategy from within-school-year, between-classes identification to within-teacher, over-time identification. This change is imposed by the

[58] To allow for comparability across different IQ thresholds, we keep the estimation sample equal by running all regressions in Appendix Figure A.6 on the sample of students with IQ below 110 (29,862 observations).

[59] For instance, Dee (2007) shows that assignment of children to a same-gender teacher improves their school achievement significantly (in terms of scores, student's engagement with the subject as well as teachers' perception of student performance). Carrell, Page, and West (2010) exploit random assignment to teachers in the U.S. Air Force Academy to document that, while males are not sensitive to the teacher's gender, having a female instructor increases females' performance in STEM and likelihood to choose a STEM-related career path (see also Mansour et al., 2018). Focusing on stereotypes, Carlana (2019) shows that teachers holding implicit negative stereotypes about female students' ability to excel in math have a negative and quantitatively significant influence on female students' performance and career choices. Similar evidence is documented by Alan, Ertac, and Mumcu (2018).

data because no teacher holds two classes in the same year.[60] The results, presented in Appendix Table A.4, show the peer effect estimates net of teachers' time-constant characteristics. We bring evidence that the results do not change when within-teacher estimations are conducted: both point estimates and significance remain very much alike to those presented in the main analysis. From these findings we surmise that teacher characteristics do not explain the gender heterogeneity documented in the main analysis.

One might be concerned that teachers adapt their (instructional) behavior depending on the presence or absence of gifted students in their class. If so, teacher fixed effects would not totally account for teachers' adaptation in behavior, teaching style, or teaching goals induced by the presence of a gifted student. Teachers in the presence of a gifted student might either adapt their teaching style towards the whole classroom or provide gifted students with personalized teaching. In the first case, teachers must weigh the interests of the gifted students with the interest of the other students and the composition of the classroom, which we control for in our estimation. In the second case, our estimates are immune to the effect of gifted students on themselves. In both cases, teachers do not know *ex ante* the gifted status of their students when they are assigned to a class, which would make any adaptation slow and more costly.

Finally, we investigate whether other students' characteristics could alternatively explain heterogeneous effects of exposure to gifted students in the classroom. As presented in Appendix Table A.5, we document that relatively young students, students with a foreign language as mother tongue, and students with school teacher of the same gender do not react differently to gifted peers. By looking at class size, we find that the impact of gifted peers is slightly larger in smaller classes. This effect is significant (5% level) and might indicate that there are dilution effect of putting one gifted student in a large class. In our main specification, we account for this dilution effect by always adding class size as a control variable. In conclusion, we are confident that these channels do not tamper with our main results and that gender is the main driving force behind our findings.

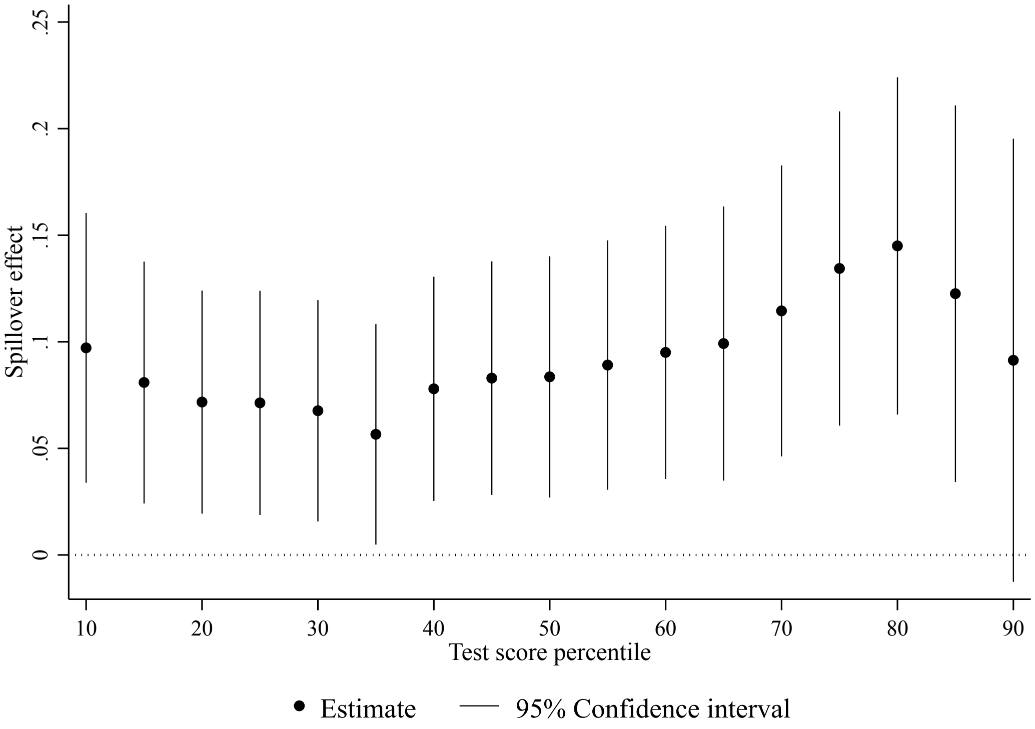
[60] In practice, we specify the error term as $\varepsilon_{icst} = \pi_s + \phi_t + \eta_{cs} + u_{icst}$, where $\eta_{cs}$ represents the teacher fixed effect ($\pi_s$ and $\phi_t$ are school and time fixed effects, respectively).

### 3.4.2 Heterogeneity and Mechanisms

Provided that we uncover no source of heterogeneity other than the gender-subject result we presented previously, what exactly drives the effect heterogeneity across gender and why does the gender gap in math achievement widen when students are exposed to gifted peers? In this section, we explore several possible explanations for the peculiar gender-subject heterogeneity in the ability peer effect proposed by the scientific literature in economics, psychology, and education science.

In a first step, we try to understand which students are affected the most by the presence of gifted classmates. We look at quantile effects of the exposure to a gifted peer. In Figure 3.2, we estimate the treatment effect of the exposure to gifted peers for classmates belonging to a given percentile of the test score distribution. The figure plots (unconditional) quantile treatment effects (following Firpo, Fortin, and Lemieux, 2009) and the respective 95%-confidence intervals for different percentiles of the achievement distribution. We find that while exposure to gifted peer has positive effects throughout the achievement distribution, students in the lower tail of the distribution do not benefit as much as students at the top. The peer effect reaches a peak around the eighth decile, indicating that high-achieving but non-gifted students react the most to the presence of gifted peers. Not only do these findings point out that mainstreaming of high ability peers in the classroom has positive peer effects for children who are academically strong – as already suggested by Card and Giuliano (2014) and Duflo, Dupas, and Kremer (2011) – but they also show that children on the lower end of the performance distribution benefit from such peers.

When we assess gender-specific impacts of being exposed to a gifted peer for children in different quantiles of the test scores distribution, we do not only find a gender penalty of having a gifted peer in math across the whole performance distribution, but also that this gender penalty appears to be significant primarily among female students at the bottom of the score distribution. Appendix Figure A.8 displays this pattern: the gender penalty is distinguishable from zero for the lowest three deciles of the test score distribution. Moreover, as suggested by our previous results, this penalty disappears in language achievement. In summary, both male and female peers in the upper tail of the achievement distribution are positively impacted by their exposure to gifted students. However, low-achieving female students seem to be relatively more negatively impacted than their low-achieving male counterparts in math. These results



*Notes:* Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Figure 3.2: Quantile treatment effect of exposure to gifted classmates

suggest that the widening of the gender gap might be driven primarily by decreased performance of low-achieving female students.

A possible – and much discussed – culprit for the reported gender difference in reaction to high ability peers is the fact that male students and female students perceive competition differently (Niederle and Vesterlund, 2011; Iriberri and Rey-Biel, 2019). On the one hand, female students have a higher tendency than male students to experience test pressure, which puts them at disadvantage in standardized tests (e.g., Gneezy, Niederle, and Rustichini, 2003; Montolio and Taberner, 2018; Saygin, 2020). On the other hand, female students shy away from competitive environments (e.g., Niederle and Vesterlund, 2007; Morin, 2015) because of lower self-confidence, lower academic self-concept, or lower willingness to compete against male students. This phenomenon is more prominent among high-ability students (Buser, Peter, and Wolter, 2017; Preckel et al., 2008).

A further mechanism that could possibly explain the heterogeneous impact of gifted peers on both male and female students is a "role model" mechanism. Gifted students may be seen as a source of inspiration by their peers and may influence their peers' achievement and motivation positively, especially if the gifted students are themselves exemplary peers. In recent years, the presence of female role models – or the lack thereof – has been shown to have a significant impact on behaviors, preferences and career choices of other female students and women, and it has emerged as a prominent explanation for the gender imbalance in STEM careers (Buckles, 2019; Avilova and Goldin, 2018; Porter and Serra, 2019).[61] In this spirit, we investigate whether female students react differently to gifted female peers than to gifted male peers (and inversely). In addition, we look at whether gifted female or male students who are exemplary students influence their female and male peers more positively than disruptive gifted students.

In Table 3.5, we decompose the effect of exposure to gifted peers into exposure to gifted female students and gifted male students and we estimate these effects on female students and male students separately. Note that we lower the IQ threshold to 115 for this analysis (one standard deviation above the mean instead of two), because the low prevalence of gifted students causes a loss of statistical power when using the 130-threshold. However, as presented previously in Figure A.6, the magnitude of the

[61] Female role models are defined as "women who can influence role aspirants' achievements, motivation, and goals by acting as behavioral models, representations of the possible, and/or inspirations" (Morgenroth, Ryan, and Peters, 2015, p. 4).

ability spillover is not greatly affected by the choice of the IQ threshold. We bring evidence that male students emulate their gifted peers irrespective of the gender of their gifted peers. As shown in Panel A of Table 3.5, coefficients of exposure to gifted male students are similar in magnitude to the coefficients of exposure to gifted female students. Interestingly, male students seem to be less affected by gifted peers when it comes to language: the size of the coefficient on language alone is twice as small as the coefficient on math. We conduct a simple F-test to show that the two coefficients are not statistically different from each other, which reinforces our conclusion that male students benefit from gifted peers irrespective of the gender of their gifted peer.

Turning our attention to female students, we find that female students react strongly to the gender of their gifted classmates. Panel B of Table 3.5 documents that female students react much less when they are exposed to gifted male students; however, they strongly react to the exposure to other gifted female students (around 10% of a standard deviation in overall test score). The only exception is in language, where female students are positively affected by gifted male students as much as by gifted female students. For math, we can conclude that female students are responsive to the gender of their gifted peer, and that they experience a positive influence in the presence of gifted female students. Once again, we stress the fact that the presence of a gifted female student in a classroom is as good as random, allowing to draw causal conclusions from our quasi-experimental setting. Not only is this finding in line with the above-mentioned literature, but it also adds valuable understanding on the importance of same-gender role models.

So far, we have implicitly assumed that all gifted students are "good peers", in the sense that they positively affect and influence their peers. However, this might not be true, especially if gifted students suffer from emotional, behavioral or social problems and disrupt the classroom. Evidence that psychological disorders such as ADHD negatively impact the academic performance, increase the tendency to underachieve and impair executive functioning of gifted individuals as much as non-gifted individuals is well-documented (Antshel et al., 2008; Brown, Reichel, and Quinlan, 2009; Mahone et al., 2002; Gomez et al., 2019).[62] It is therefore important to distinguish between gifted children who, by their disruptive behavior, are less likely to be seen as exemplary stu-

[62] Gifted students who suffer from other mental health disorders are also referred to as "twice-exceptional." There is a debate in the literature in psychology on whether gifted individuals are relatively more likely as the general population to develop psychological disorders (such as ADHD, anxiety, and mood disorders) or depression. Some researchers suggest that gifted individuals are more inclined to developing such disorders and defend an "overexcitability perspective," such as Karpinski

| | (1)<br>Composite<br>test score | (2)<br>Math<br>test score | (3)<br>Language<br>test score |
|---|---|---|---|
| | A. Male students | | |
| Exposure to gifted male classmates | 0.058** | 0.075*** | 0.025 |
| | (0.027) | (0.027) | (0.026) |
| Exposure to gifted female classmates | 0.085** | 0.094*** | 0.052 |
| | (0.034) | (0.032) | (0.032) |
| *F-test for equality of coefficients (p-value)* | *0.535* | *0.635* | *0.510* |
| | B. Female students | | |
| Exposure to gifted male classmates | 0.046 | 0.029 | 0.051* |
| | (0.029) | (0.028) | (0.027) |
| Exposure to gifted female classmates | 0.105*** | 0.125*** | 0.055** |
| | (0.029) | (0.030) | (0.026) |
| *F-test for equality of coefficients (p-value)* | *0.152* | *0.018*** | *0.921* |
| Individual characteristics | Yes | Yes | Yes |
| Classroom characteristics | Yes | Yes | Yes |
| School-by-year FE | Yes | Yes | Yes |

*Notes:* *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$. Standard errors, shown in parentheses, are clustered at the classroom level. In panel A, $N = 14,274$; in panel B, $N = 16,203$. Individual characteristics include gender, native German speaker, and age at test. Classroom characteristics include class size, share of females, share of native German speakers, and average age at test. Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Table 3.5: Spillovers by gender of the gifted

dents by their peers, and gifted students who are more likely to foster productivity and generate positive externalities in the classroom.

Having information on the psychological profile of each child sent to the School Psychological Service, we identify gifted students who have been also referred for exhibiting behavioral, emotional or social difficulties (12.3% of all gifted students, see Table 3.1). Children with such difficulties are usually referred to the SPS for disrupting the classroom or for showing mental or emotional problems. In terms of school performance, we find no visible difference in the distribution of test scores for gifted children with and without emotional or behavioral disorders (see Figure A.10).

As presented in Table 3.6, we find that gifted children with emotional issues have a statistically insignificant influence on their peers, despite the negative sign of the coefficient. When looking at the effect of the presence of gifted children without emotional difficulties, the effect is not only significantly positive, but also 25% larger than the effect of exposure to all gifted children together. This suggests that only gifted peers who are not disturbing the classroom environment have positive impact on their peers; disruptive high ability peers at best have no impact and at worst put their peers' performance in jeopardy. This finding can be made even more salient when we analyze which students in the test score distribution are affected. Appendix Figure A.9 shows, on the one hand, that gifted children with emotional issues do not affect their peers (although there are significant negative effects around the median). On the other hand, gifted children without disruptive tendencies positively inspire all their peers towards better performance, with the effect being slightly larger among high achievers – as in our main analysis.

In addition, we find that female students are more sensitive to the presence of disruptive gifted students than male students. Whereas male students are not affected by the presence of gifted peers with behavioral, emotional, or social issues, female students react strongly and negatively. Column 4 of Table 3.6 reports the interaction effects of being a female student with the exposure to an emotionally unstable gifted peer. Female students perform dramatically less well when they are put in the same class as an emotionally unstable gifted student (about 28% of a standard deviation less than male students), while male students do not react at all to the same gifted peers.

et al. (2018) who survey adult members of Mensa (the largest and oldest high IQ society in the world, which is open to people who score at the 98th percentile or higher on a standardized, supervised IQ or other approved intelligence test), whereas others argue that the prevalence of psychological disorders is not higher among the gifted (e.g., Peyre et al., 2016).

| | (1)<br>Composite<br>test score<br>(all students) | (2)<br>Composite<br>test score<br>(male students) | (3)<br>Composite<br>test score<br>(female students) | (4)<br>Composite<br>test score<br>(all students) |
|---|---|---|---|---|
| Exposure to gifted with difficulties | -0.045 | -0.022 | -0.076 | 0.009 |
| | (0.047) | (0.057) | (0.061) | (0.053) |
| Exposure to gifted without difficulties | 0.123*** | 0.166*** | 0.075** | 0.139*** |
| | (0.024) | (0.029) | (0.031) | (0.027) |
| Female | -0.181*** | | | -0.170*** |
| | (0.015) | | | (0.016) |
| (Exposure to gifted with difficulties)* | | | | -0.110** |
| (Female) | | | | (0.055) |
| (Exposure to gifted without difficulties)* | | | | -0.029 |
| (Female) | | | | (0.028) |
| Individual characteristics | Yes | Yes | Yes | Yes |
| Classroom characteristics | Yes | Yes | Yes | Yes |
| School-by-year FE | Yes | Yes | Yes | Yes |
| Observations | 31,187 | 14,709 | 16,478 | 31,187 |

*Notes:* *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$. Standard errors, shown in parentheses, are clustered at the classroom level. Individual characteristics include gender, native German speaker, and age at test. Classroom characteristics include class size, share of females, share of native German speakers, and average age at test. Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Table 3.6: Exposure to gifted with and without social, behavioral, or emotional difficulties

However, both male students and female students react positively to the presence of gifted peers who do not have emotional issues, and there is no statistically significant gender difference in this respect.[63]

In sum, our results strongly suggest that both the quality of high-ability spillovers and the sensitivity to disruption play a role in the widening of the gender gap for students exposed to gifted peers. In fact, being exposed to gifted classmates is no guarantee of improved school performance. Instead, the way high ability peers behave in class is a critical factor for the development of virtuous spillover effects, especially when it comes to spillovers on female students. Our empirical findings are consistent with both the role-model interpretation and previous research on the detrimental effects of disruptive students (e.g., Carrell, Hoekstra, and Kuka, 2018).

[63] The same patterns discussed here hold if we perform the analysis for math and language separately.

### 3.4.3 Trajectories after Compulsory Education

Although the results presented here already have far-reaching educational policy consequences, the question remains whether the impact that gifted students have on their peers is limited to academic achievement during the time spent together, or whether it has longer-term consequences beyond that time? Linking our data with administrative data of the educational system allows us to follow the educational career of around 75% the students beyond the period of compulsory schooling. Attrition is almost exclusively due to individuals not having yet completed compulsory education (i.e., the last two school cohorts in the data). Specifically, we are interested in whether the presence of a talented peer in the class influences the peers' educational choices. In Switzerland, tracking into a vocational track or academic track occurs after compulsory education, when students are about 16 years old. A minority (less than twenty percent) of the students opt for the selective academic track (baccalaureate schools) and the majority (two third of a cohort) choses a vocational track, which in the region under observation is usually offered in the form of apprenticeships.[64]

By looking at whether students choose an academic track, a vocational track, or no post-compulsory education at all, we examine whether being exposed to gifted students has longer-run effects. Results are presented in Table 3.7, divided into three binary outcomes as follows: no post-compulsory education started, vocational track started, and academic track started. In each regression, the reference category is always the other two trajectories combined, in order to avoid conditioning on downstream outcomes.

We find that being exposed to gifted classmates in secondary school significantly increases the likelihood of choosing the academic track. Interestingly, this effect is entirely driven by male students who enter the academic track instead of the vocational one. No effect is found for female students, except for a small and marginally significant negative effect on the probability of starting vocational education. In general, we find that exposure to gifted peers during secondary school does not change the probability of pursuing any post-compulsory education degree. This result is expected, given our previous findings that low-achieving students are less affected than high-achievers by the presence of gifted students.

[64] A detailed overview of the Swiss education system can be found here: `http://www.edk.ch/dyn/11586.php`.

| | (1)<br>No post-compulsory<br>education started | (2)<br>Academic track<br>started | (3)<br>Vocational track<br>started | (4)<br>Vocational occupation<br>in STEM field |
|---|---|---|---|---|
| *Panel A. Full sample* | | | | |
| Exposure to gifted classmates | 0.003 | 0.029*** | -0.032*** | 0.029*** |
| | (0.006) | (0.010) | (0.010) | (0.009) |
| *Panel B. Male students* | | | | |
| Exposure to gifted classmates | -0.004 | 0.041*** | -0.037*** | 0.054*** |
| | (0.008) | (0.012) | (0.012) | (0.016) |
| *Panel C. Female students* | | | | |
| Exposure to gifted classmates | 0.013 | 0.015 | -0.027** | 0.001 |
| | (0.008) | (0.013) | (0.013) | (0.006) |
| Individual characteristics | Yes | Yes | Yes | Yes |
| Classroom characteristics | Yes | Yes | Yes | Yes |
| School-by-year FE | Yes | Yes | Yes | Yes |

*Notes:* *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$. Standard errors, shown in parentheses, are clustered at the classroom level. Number of observations for columns (1)-(2)-(3): in panel A, $N = 25,353$; in panel B, $N = 12,036$; in panel C, $N = 13,317$. Number of observations for column (4): in panel A, $N = 14,205$; in panel B, $N = 7,241$; in panel C, $N = 6,964$. Individual characteristics include gender, native German speaker, and year of birth. Classroom characteristics include class size, share of females, share of native German speakers, and average age at test. Data are from the School Psychological Service St. Gallen, the Ministry of Education of the canton of St. Gallen, and the Stellwerk test service provider.

Table 3.7: Post-compulsory education trajectories

We complement the results by looking at whether exposure to gifted peers affects the choice of vocational career for those students starting a Vocational Education and Training (VET). We use data from on the exact values of the apprenticeship requirement level ratings.[65] These ratings are made by experts who assess the requirement content of each VET occupation along the following dimensions: math, science, first language, and foreign language.[66] We code a VET occupation as "STEM-intensive" when its curriculum content in math and science belongs to the top quartile of the math and science content distribution among all occupations. When focusing on vocational education, we find that exposure to gifted students in the classroom increases the enrollment in STEM-intensive occupations among students who choose a vocational career. Again, we find that only male students are the ones affected by the exposure to gifted peers when choosing their vocational occupation. Our results are in line with recent evidence that peer characteristics during adolescence influence important career decisions (Anelli and Peri, 2017; Carrell, Hoekstra, and Kuka, 2018; Black, Devereaux, and Salvanes, 2013). They also highlight the importance of peer quality on the decision to pursue STEM-related careers among females, as highlighted by Mouganie and Wang (2020) and Card and Payne (2017).

However, unlike Mouganie and Wang (2020) who examine high school students in China, we do not find a negative effect of high-ability male students on women's likelihood to choose a science track during high school. One obvious explanation for this discrepancy is the difference in the institutional setting. Another difference is that we focus on gifted students, a population with higher ability on average but with heterogeneous school achievement. Thus, the results of the present paper constitute an important complement to the findings for high-ability students in China.

## 3.5 Conclusion

The present study sheds light on the relevance of gifted students and their heterogeneous spillovers effects in the classroom. Heterogeneity is observable in at least three dimensions. First, not all gifted students impact their peers in the same manner, but the impact depends on the gender of the gifted student and also whether gifted

[65] Attrition on the data on occupational profiles is almost exclusively due to the impossibility to merge our data with the corresponding apprenticeship requirement level ratings (see Appendix Table A.2).

[66] These data are the product of an initiative by the Swiss Cantonal Ministers of Education, with financial support of the State Secretariat for Education, Research, and Innovation (SERI).

students show behavioral problems or not. Second, not all peers are affected in the same way, but effects differ by gender and ability of the peers and third, peers are not affected in the same way in all subjects. We find that while male students benefit from the presence of gifted peers in all subjects regardless of their gender, female students benefit primarily from the presence of gifted female students. The nature of our data allows us to test a number of potential mechanisms. We show that neither teachers nor classroom composition are responsible for driving the heterogeneity in the ability spillovers. Instead, we present suggestive evidence consistent with the hypothesis that academic role models and classroom behavior are important determinants of the gender gap in math.

Moreover, our results suggest that exposure to gifted students has powerful, lasting effects on career choices and post-compulsory education. The presence of gifted students in the classroom is a catalyst for pursuing an academic track or a STEM-intensive vocational training. However, this catalytic effect is found to perpetuate (and even deepen) the gender gap in the likelihood of choosing occupations that require higher STEM skills. Indeed, men react to the presence of gifted peers by taking up STEM-intensive occupations, whereas women do not. With the disclaimer that our results measure the peer effect of *identified* gifted students, we corroborate existing evidence that social environment affects women's STEM career choices as early as high school.

In general, we find that gifted students are influential in fostering emulation and impacting positively the academic achievement and the career choices of their peers. They are therefore fundamental forces in the classroom production function that should not be ignored in designing successful educational policies, especially when considering whether gifted students should be segregated in more "elite" schools or pull-out programs. We also show that giftedness alone is no guarantee for positive externalities: gifted students who were diagnosed with socio-emotional problems generate null-to-negative spillovers on their peers.

In terms of classroom composition policies in an inclusive education system, the results of the present paper offer two major insights. First, it is desirable to evenly spread gifted students without behavioral problems throughout classrooms. This reassignment scheme would ensure that non-gifted students have increased chances of being exposed to non-disruptive gifted students and benefit from positive learning spillovers. Second, reassignment policies should allocate disruptive gifted students randomly to classrooms, because doing so would ensure that non-gifted students have random chances of being exposed to disruptive gifted students, which is in line with the principle of equality of chances. These recommendations come, however, with one important caveat. Our models estimate the effect of gifted peers on non-gifted students, but they do not provide us with the effect of gifted peers on fellow gifted students. As such, we miss an important ingredient of sound reallocation policies, i.e., how much gifted students themselves benefit – or not – from each other and from their non-gifted peers. Therefore, further research calls for a better understanding of optimal classroom allocation, possibly balancing the positive spillovers from high-ability students with negative spillovers from disruptive students.

# 4 | The Earth is Not Flat:
# A New World of High-Dimensional Peer Effects[*]

Aurélien Sallin (University of St. Gallen)
Simone Balestra (University of St. Gallen and CESifo)

*We develop a novel machine learning empirical framework to better understand peer effects in the classroom. This approach 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 performance, (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 own academic performance, and that the strongest peer effects are generated by students with special needs, low-achieving students, and male students. Second, we show that classroom peer effects 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 ones that reach maximal aggregated school performance.*

**Keywords**: peer effects, high dimensionality, machine learning, classroom composition
**JEL Classification**: C31, H75, I21, I28

[*] We are grateful to Aline Bütikofer, Beatrix Eugster, Daniel Goller, Michael Lechner, Jana Mareckova, 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.

## 4.1 Introduction

In the past 15 years, the literature on peer effects in the school environment has grown exponentially (e.g., Balestra, Sallin, and Wolter, forthcoming; Brenøe and Zölitz, 2020; Carrell, Hoekstra, and Kuka, 2018; Angrist, 2014; Burke and Sass, 2013; Black, Devereaux, and Salvanes, 2013; Lavy, Silva, and Weinhardt, 2012; Lavy and Schlosser, 2011a; Bifulco, Fletcher, and Ross, 2011; Carrell and Hoekstra, 2010), and we now have a better understanding of the main driving forces at play in the classroom. However, most of the classroom peer effects[67] 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 peer effects have heterogeneous impacts on individuals with different characteristics, and they do not incorporate the reality that there are *at least* as many peer effects 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 peer effects.

In this paper, we provide a more comprehensive analysis of educational peer effects by looking at the systematic interaction of peers and their many characteristics. We start with the assumption that "true" peer 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, peer 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.

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

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 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 and Imbens, 2019; Athey, 2019; Mullainathan and Spiess, 2017), but no study so far has used such methods to address the problem of educational peer 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 educational needs (SEN) (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 non-native 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. The fact that interactions 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 peer 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., Bietenbeck, 2020; Balestra, Eugster, and Liebert, forthcoming). 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 peer effects. 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 peer 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, peer effects from gender are virtually nonexistent. Similarly, we deliver a different picture of peer effects 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 peer effects are not constant.[68] 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 peer 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 Isphord-

[68] 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.

ing 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 peer 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 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 many 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 making classrooms more segregated in students' characteristics does not improve the aggregate academic performance of students.

## 4.2 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.[69] 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[70], 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 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.[71]

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 iden-

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

[70] 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.

[71] 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.

tifier, 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.[72] 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).[73]

We merged the two administrative data sets using anonymous student identifiers 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 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).

Table 4.1 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 that 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

[72] 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.

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

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 students, 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).

| | 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 4.1: Summary statistics

## 4.3 Setting

### 4.3.1 Empirical setting

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

$$y_{ic} = \alpha + \beta T_{ic} + \gamma \overline{T}_{(-i)c} + \delta \mathbf{X}_{ic} + \mu_c + \epsilon_{ic}, \tag{4.1}$$

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 $\epsilon_{ic}$ 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 $\overline{T}_{(-i)c} = \overline{t}'_{(-i)c}$ and $\overline{T}_{(-i)c} = \overline{t}''_{(-i)c}$ 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 (4.1) 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 4.3.3 and Section 4.3.4 below), we now turn our attention to the identification strategy.

### 4.3.2 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 students' characteristics 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., Black, Devereaux, and Salvanes, 2013; Angrist and Lang, 2004; Bifulco, Fletcher, and Ross, 2011; Brenøe and Zölitz, 2020; Carrell and Hoekstra, 2010; Carrell, Hoekstra, and Kuka, 2018; Lavy, Silva, and Weinhardt, 2012; Lavy and Schlosser, 2011a). However, peer 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, our 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 4.2, 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.

| | 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 two-sample $t$-test, and p-values are reported.

Table 4.2: Variation in cohort or classroom composition measures and randomization checks

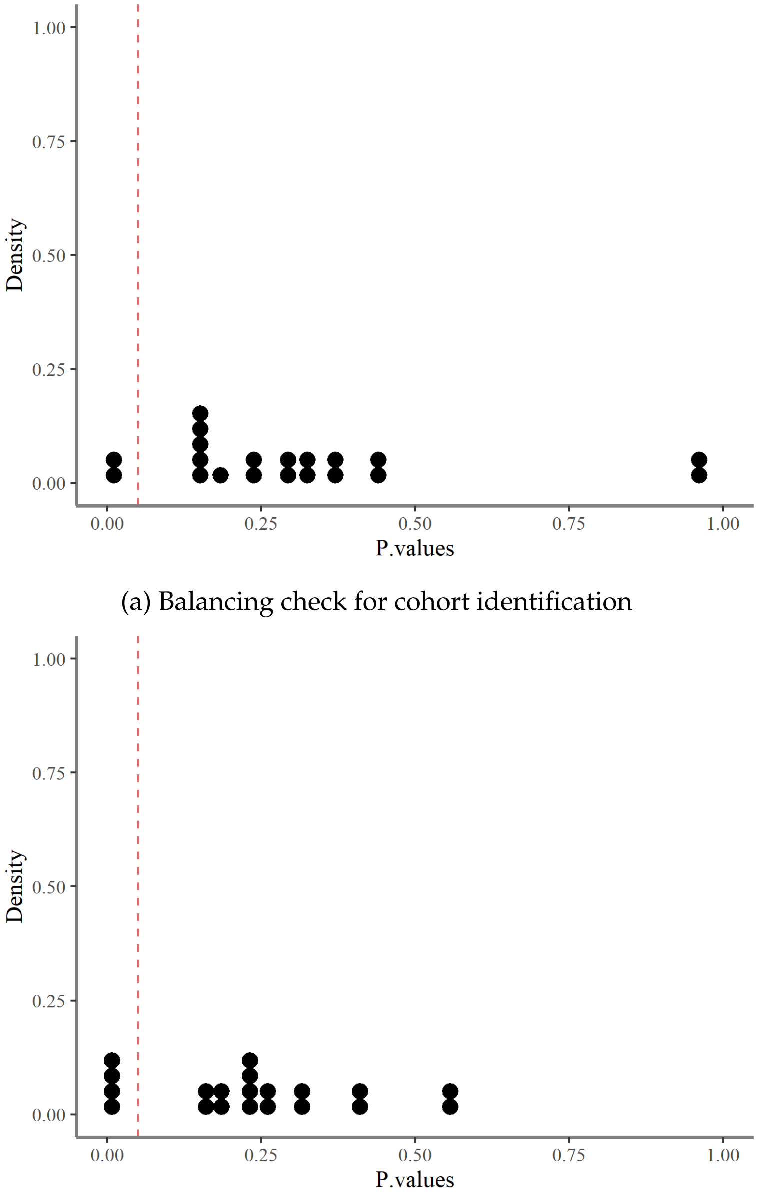
To detect potential selection into classrooms or into cohorts, we conduct further balancing checks by testing whether each peer effect variable (the mean of LoO variables) predicts individual baseline characteristics (gender, native speaker, relative age, special needs, and giftedness) conditional on 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.

The distribution of p-values for the balancing tests across classrooms is presented in Figure 4.2a. 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 our estimates "control" for both individual and cell characteristics.

### 4.3.3 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 know which variables are crucial determinants of aca-



(a) Balancing check for cohort identification

(b) Balancing check for 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 5×5 - 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 4.1: Balancing checks for cohort and classroom identification

demic 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 practically unusable.[74]

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:

$$\hat{\pi}^\lambda_j = \frac{1}{B} \sum_{b=1}^{B} I_{j \in \hat{S}^b_\lambda}, \tag{4.2}$$

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

[74] 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.

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 $\hat{S}_{stable} = \{j : \max_{\lambda \in \Lambda} \hat{\pi}_j^\lambda \geq \pi^*\}$.[75] 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 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.[76]

### 4.3.4 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 (4.1): 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 correspond-

[75] 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).

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

ing LoO proportions have the same length. We therefore want to flexibly estimate the following function:

$$y_{ic} = g(\mathbf{T}_{ic}, \overline{\mathbf{T}}_{(-i)c}, \mathbf{X}_{ic}, \zeta(\mathbf{T}_{ic}, \overline{\mathbf{T}}_{(-i)c}, \mathbf{X}_{ic})), \tag{4.3}$$

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 estimand, which conceptually corresponds to the coefficient $\gamma$ in Equation (4.1):

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

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 $\overline{t}_{c,k}$, of $\overline{t'}_{c,k}$, as well as with all the other variables $\overline{t}_{c,k}$ 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 $l$:

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

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.[77] In step (4), 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.[78]

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

[77] 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.

[78] 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.

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.

## 4.4 Results

### 4.4.1 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 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.[79]

[79] 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

**Results of the stable selection.** We present results of the stable selection exercise at the cohort level in Figure 4.3 without interactions, Figure B.1 with interactions, and Figure 4.5 of the Appendix for fully interacted types. For the classroom level analysis, we present results in Figure 4.4 without interactions, Figure B.2 with interactions, and Figure 4.6 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 $\lambda$.[80] 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 the others are held constant.[81]

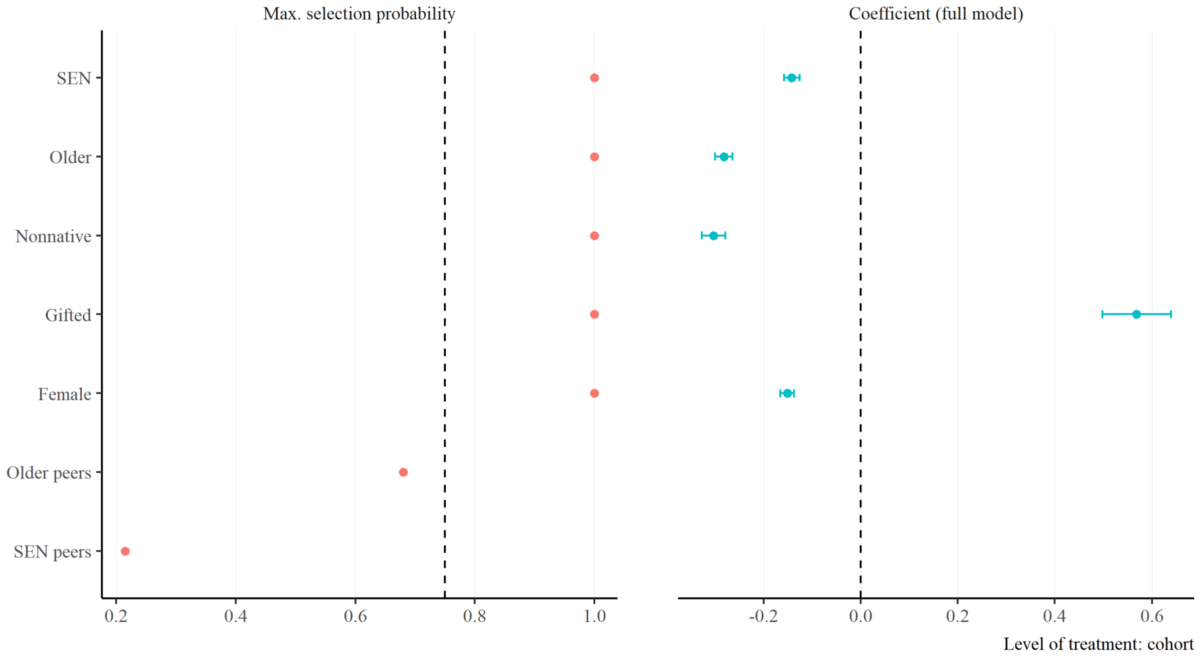
The analysis without interactions in Figures 4.3 and 4.4 reveals insightful results. At the cohort level, only the five own types are selected and no peer effects are present. This means that peers' influence at the cohort level does not predict academic performance. At the classroom level, four out 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).

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

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.

[80] 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.

[81] 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.

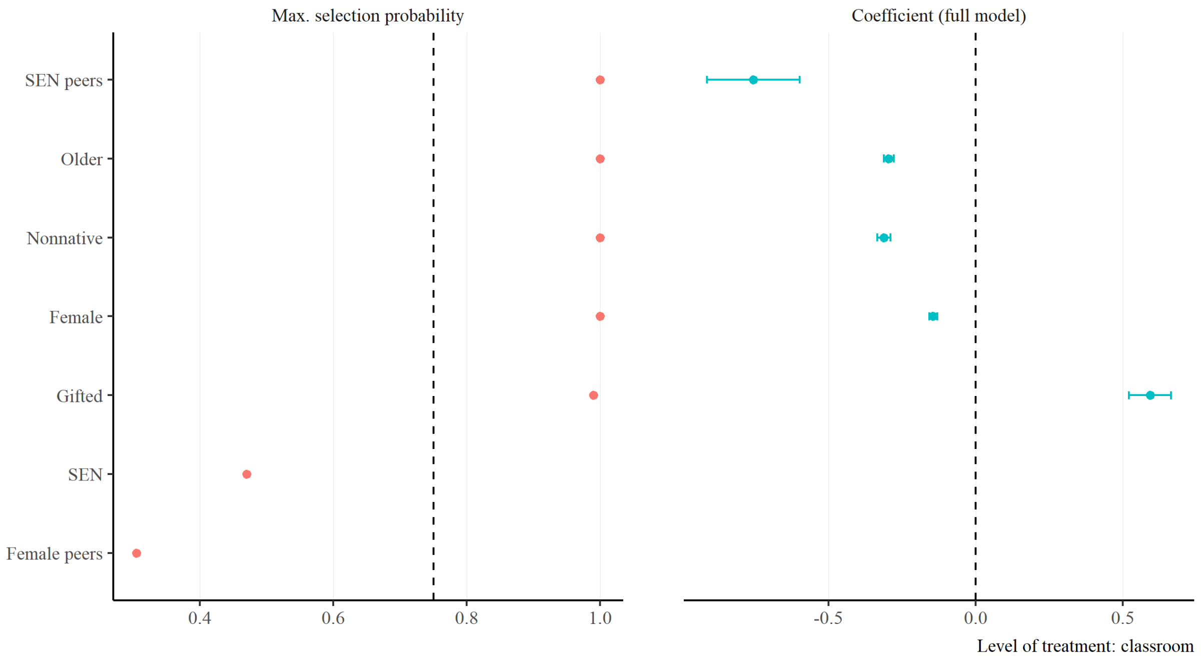


*Notes:* the left panel of this graph reports the probability for a variable to be selected. The learning algorithm is stable selection with hierarchical group lasso on 200 folds of size $n/2$. Selected variables are variable selected with a probability higher than 0.75 (dashed line). The × indicates 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 4.3.3. Effect sizes and confidence intervals are presented for information only and must be interpreted with caution.

Figure 4.3: Stability selection and effect size at the cohort level without interactions

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 C.2 (Figures B.1 and B.2), the effect of older peers and the effect of peers with SEN are the dominating peer effects at

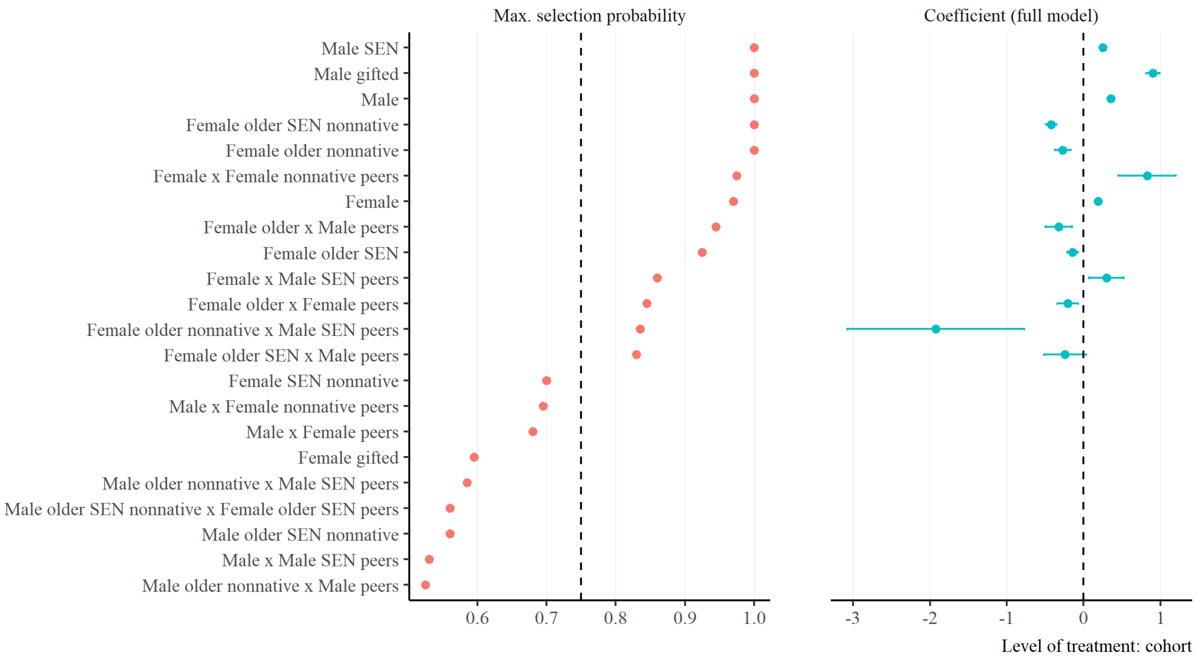


*Notes:* the left panel of this graph reports the probability for a variable to be selected. The learning algorithm is stable selection with hierarchical group lasso on 200 folds of size $n/2$. Selected variables are variable selected with a probability higher than 0.75 (dashed line). The × indicates 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 4.3.3. Effect sizes and confidence intervals are presented for information only and must be interpreted with caution.

Figure 4.4: Stability selection and effect size at the classroom level without interactions

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 effects 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. 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



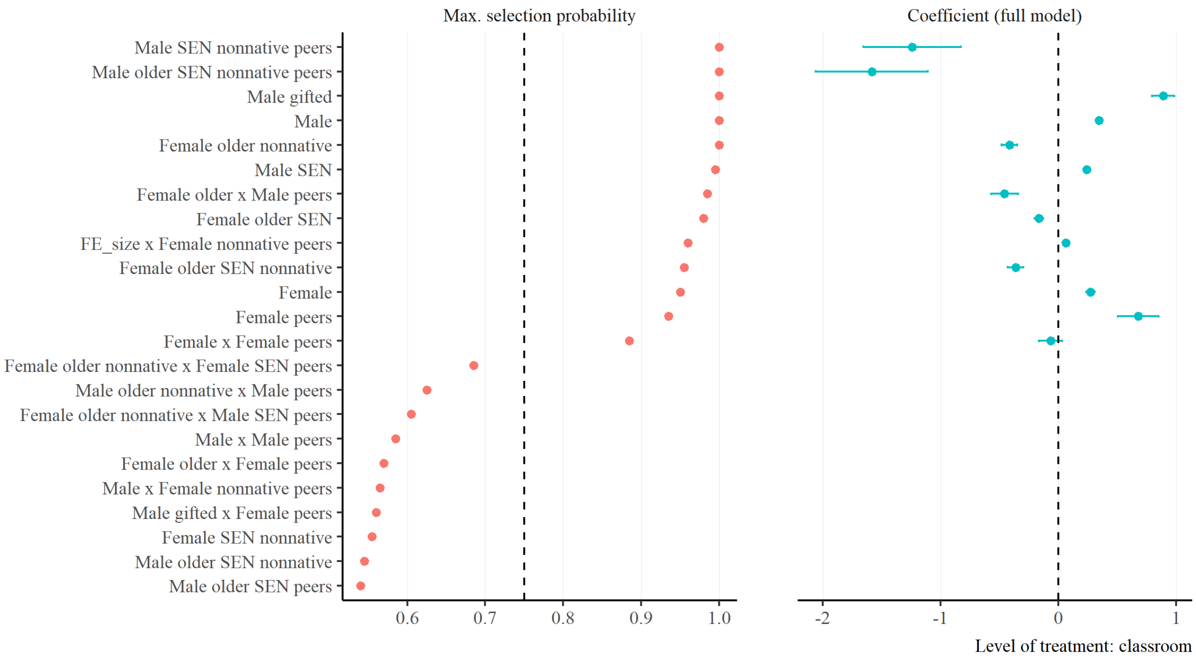
*Notes:* the left panel of this graph reports the probability for a variable to be selected. The learning algorithm is stable selection with hierarchical group lasso on 200 folds of size $n/2$. Selected variables are variable selected with a probability higher than 0.75 (dashed line). The × indicates 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 4.3.3. Effect sizes and confidence intervals are presented for information only and must be interpreted with caution.

Figure 4.5: Stability selection and effect size at the cohort level with fully interacted types

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.

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 4.5 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.[82] 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).

[82] Note that the variables "male" or "female" correspond to male and female students who are not



*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 × 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 4.3.3. Effect sizes and confidence intervals are presented for information only and must be interpreted with caution.

Figure 4.6: Stability selection and effect size at the classroom level with fully interacted types

At the classroom level, Figure 4.6 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., Bietenbeck, 2020; Carrell, Hoekstra, and Kuka, 2018; Balestra, Eugster, and Liebert, forthcoming; Bifulco, Fletcher, and Ross, 2011; Hanushek, Kain, and Rivkin, 2009; Figlio, 2007). In addition, female peers with no other observed characteristics generate positive peer effects, whereas male peers with no other observed characteristics generate positive peer effects 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 (2011a). 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. Fi-

older, not nonnative, not gifted, and without SEN.

nally, spillover from female peers 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 peer effects interact in school groups. It shows that peer 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 peer effects. Finally, it shows that effects and peer effects are most likely heterogeneous across types, and depend heavily on the group composition.

### 4.4.2 High-Dimensional Models

We estimate high-dimensional models using ML algorithm as presented in Section 4.3.4, 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.[83]

[83] 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

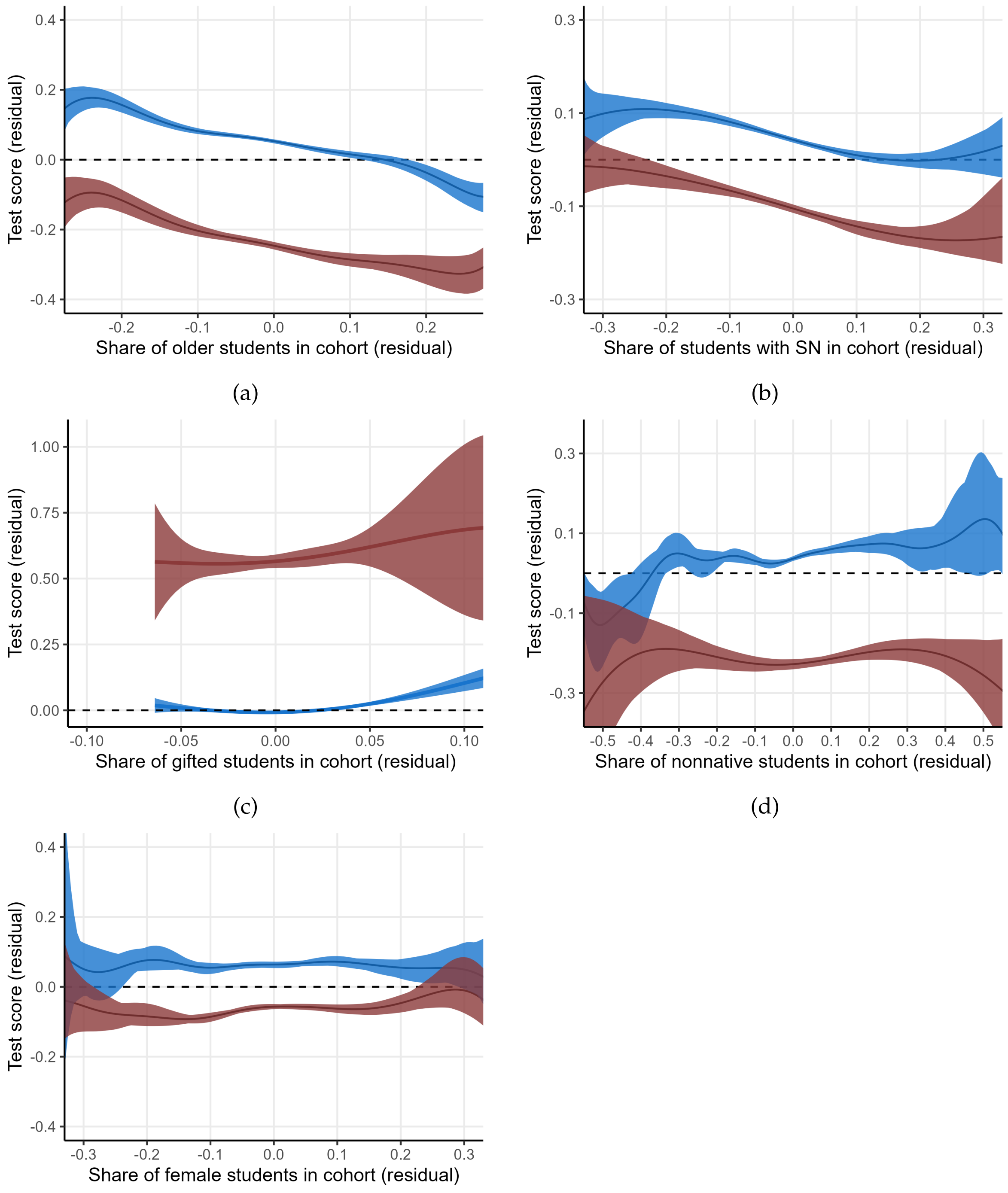
**Results at the cohort level.** We first discuss the high-dimensional functions for the cohort identification strategy, which are presented in Figure 4.7. Each subgraph presents the effect of peers of a given type (for instance, share of female peers) both on peers of the same type in red (female students), and on peers of the other type in blue (male students).

Figure 4.7a 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 4.7b. The representation of peer effects for gifted students in Figure 4.7c 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 affects neither 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.

**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 4.9a 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

*M* predictions, and the lower confidence interval across the *M* predictions. This explains the peculiar shape of confidence intervals in reported figures.



(a)

(b)

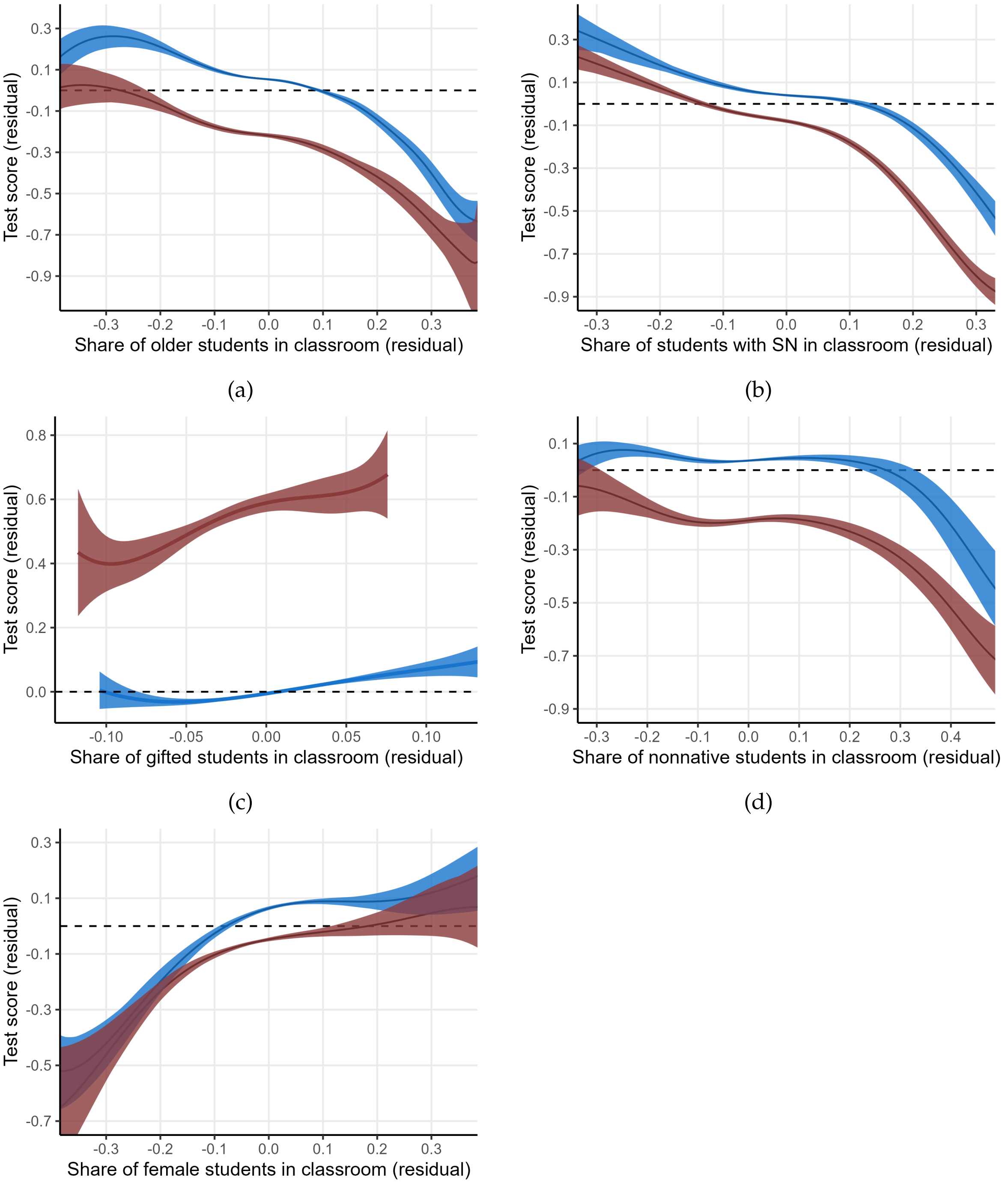
(c)

(d)

(e)

*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. For instance, if the effect of the proportion of nonnative students is investigated, the red line gives the effect of nonnative peers on nonnatives, whereas the blue line gives the effect of nonnative peers on natives. 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.

Figure 4.7: High-dimensional peer effects at the cohort level



(a)

(b)

(c)

(d)

(e)

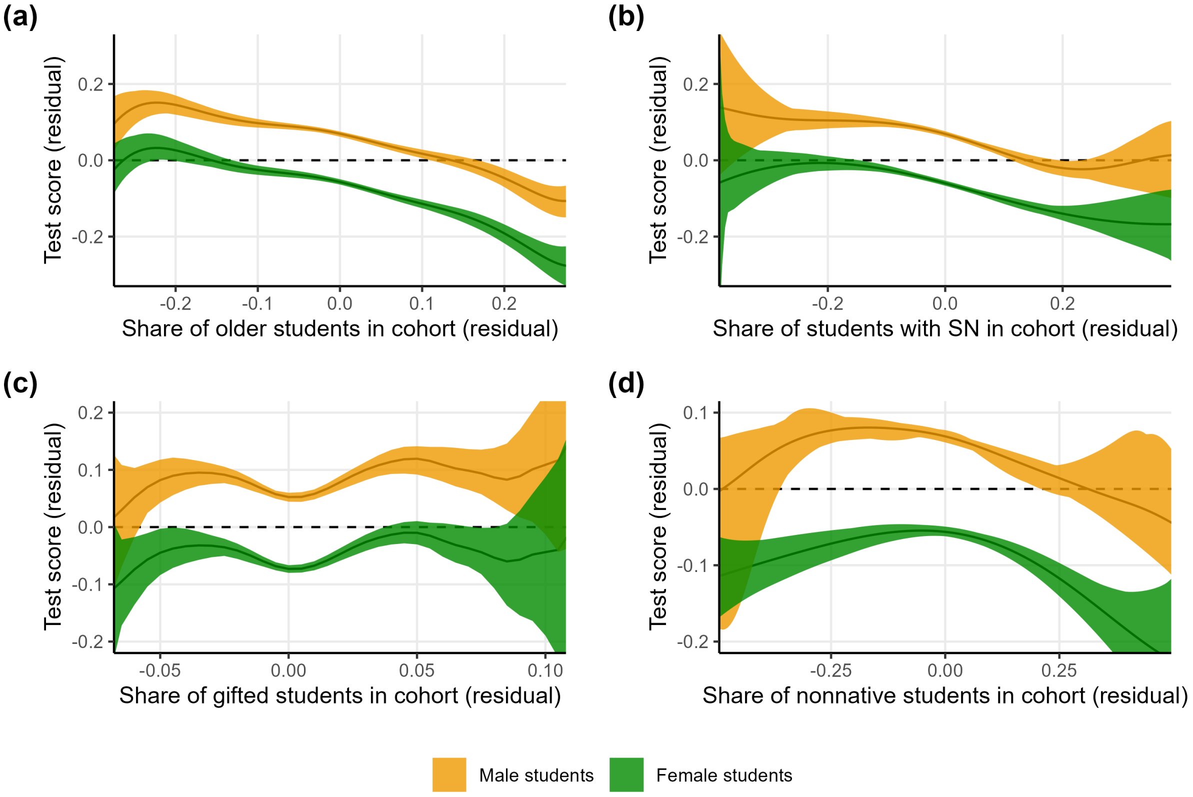
*Notes:* This figure displays the effect of the share of types of interest per classroom on the standardized test scores. Test scores and shares are demeaned from their school-track-year averages. Heterogeneity across own types is depicted in different colors: the own effect is given in red, and the effect for the other category is given in blue. For instance, if the effect of the proportion of nonnative students is investigated, the red line gives the effect of nonnative peers on nonnatives, whereas the blue line gives the effect of nonnative peers on natives. 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 4.9: High-dimensional peer effects at the classroom level

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 20 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 peer effects in classrooms rather than cohorts.

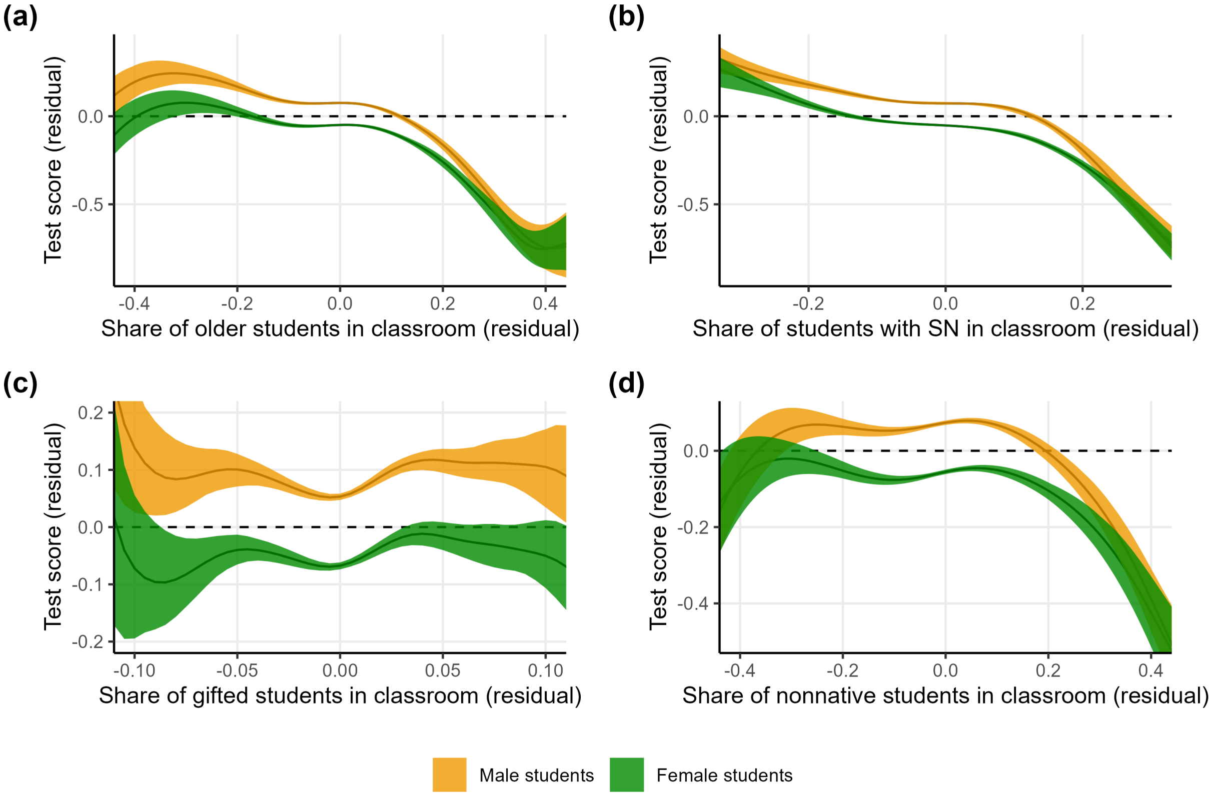
**Heterogeneity across gender** In an additional step, we investigate heterogeneities of peer effects across gender. Many studies investigate heterogeneous peer effects across gender to understand, for example, the gender gap in school performance (Fryer and Levitt, 2010), career choices (e.g., Brenøe 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 4.11 and 4.12.

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.



*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 4.11: High-dimensional peer effects: gender heterogeneity at the cohort 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.

Figure 4.12: High-dimensional peer effects: gender heterogeneity at the classroom level

| | 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 4.3: Optimal classroom and cohort environments

**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 4.3 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 result corroborates the extensive literature on class size (Lazear, 2001).

## 4.5 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 peer effects 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.3.

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*:[84]

$$\widehat{ACE} = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{y}_{ic}^{\text{segr.}} - \hat{y}_{ic}^{\text{random}}\right], \tag{4.4}$$

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 non-native students. The *CACE* is:

$$\widehat{CACE} = \frac{1}{\sum_{i=1}^{N}\mathbf{1}(T_{ic} = t_c)}\sum_{i=1}^{N}\mathbf{1}(T_{ic} = t_c)\left[\hat{y}_{ic}^{\text{segr.}} - \hat{y}_{ic}^{\text{random}}\right]. \tag{4.5}$$

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

[84] 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.

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 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 4.4. 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 classroom 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 more 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.

| | Randomized regime<br>$\frac{1}{N}\sum_{i=1}^{N}\left[\hat{Y}_{ic}^{\text{random}}\right]$ | Manipulated regime<br>$\frac{1}{N}\sum_{i=1}^{N}\left[\hat{Y}_{ic}^{\text{segr.}}\right]$ | Difference<br>*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 $^{***}$ |

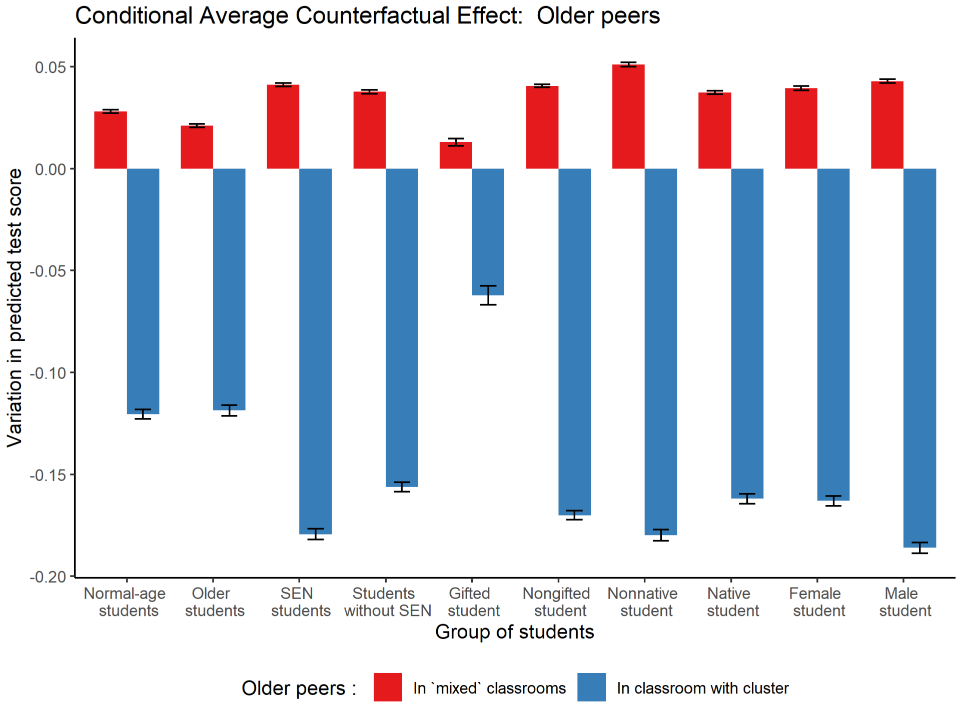
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.001$

*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 of a given type from all classrooms but one, and creates a cluster of students of the given type 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.

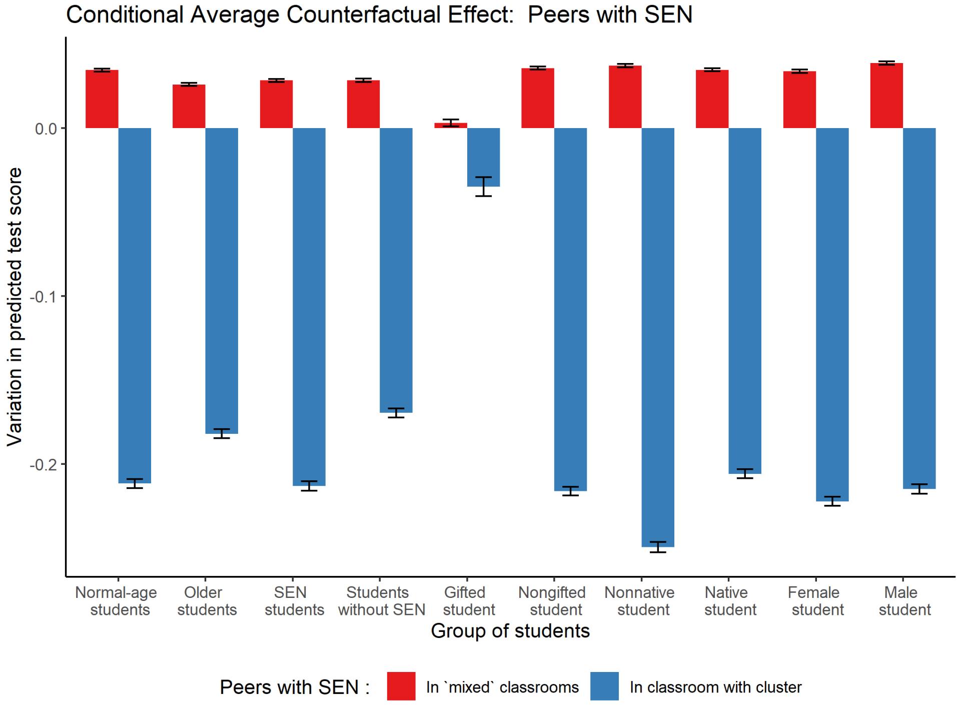
Table 4.4: Comparison of randomized and counterfactual manipulated allocation

**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 4.13, 4.15, and 4.17 present the gains and losses for students of all categories when one older student (Figure 4.13a.), when one student with SEN (Figure 4.13b.), when one nonnative student (Figure 4.15a.), when one female student (Figure 4.15b.), and when one gifted student (Figure 4.17) 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 .

What happens when we cluster students of a given type into one classroom? For all negative peer effects (peer effects from nonnative students, older students with



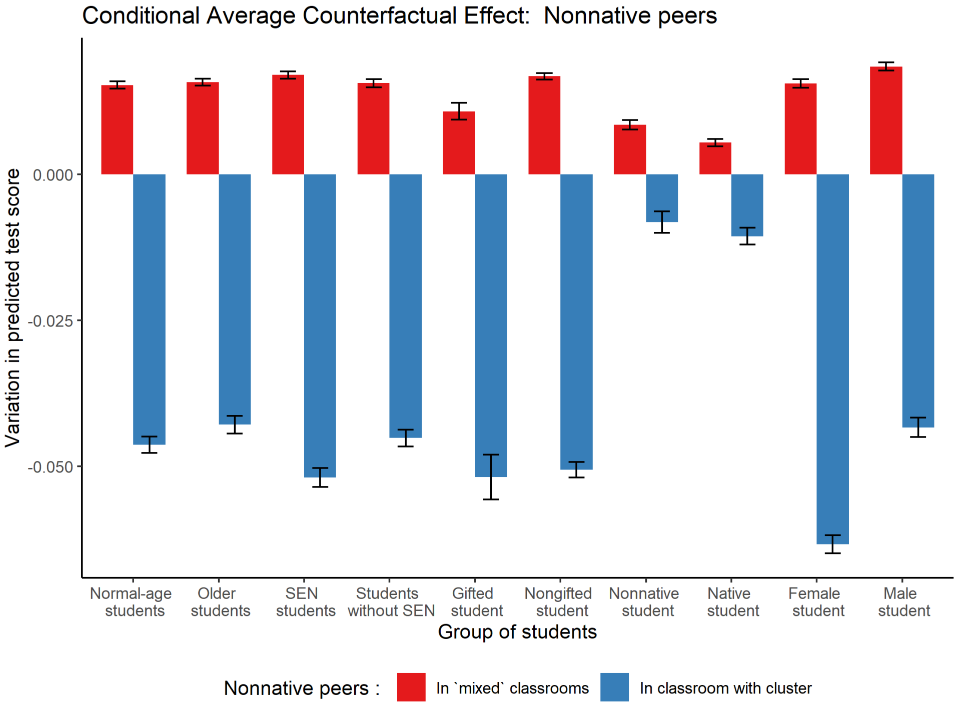
(a)



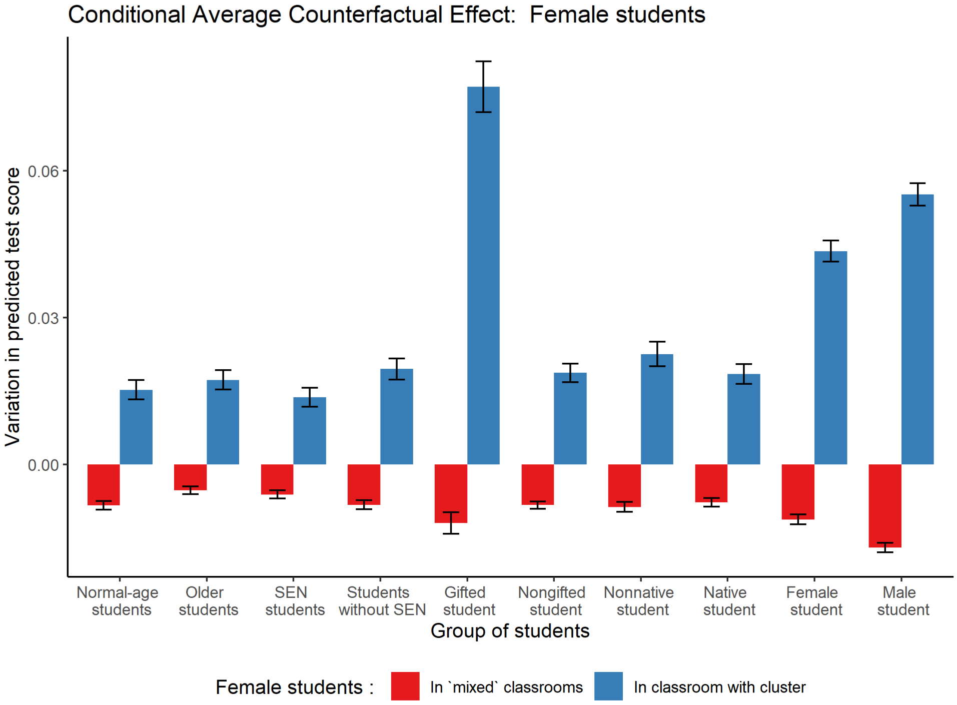
(b)

*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).

Figure 4.13: CACE: clustering of older peers and peers with SEN



(a)



(b)

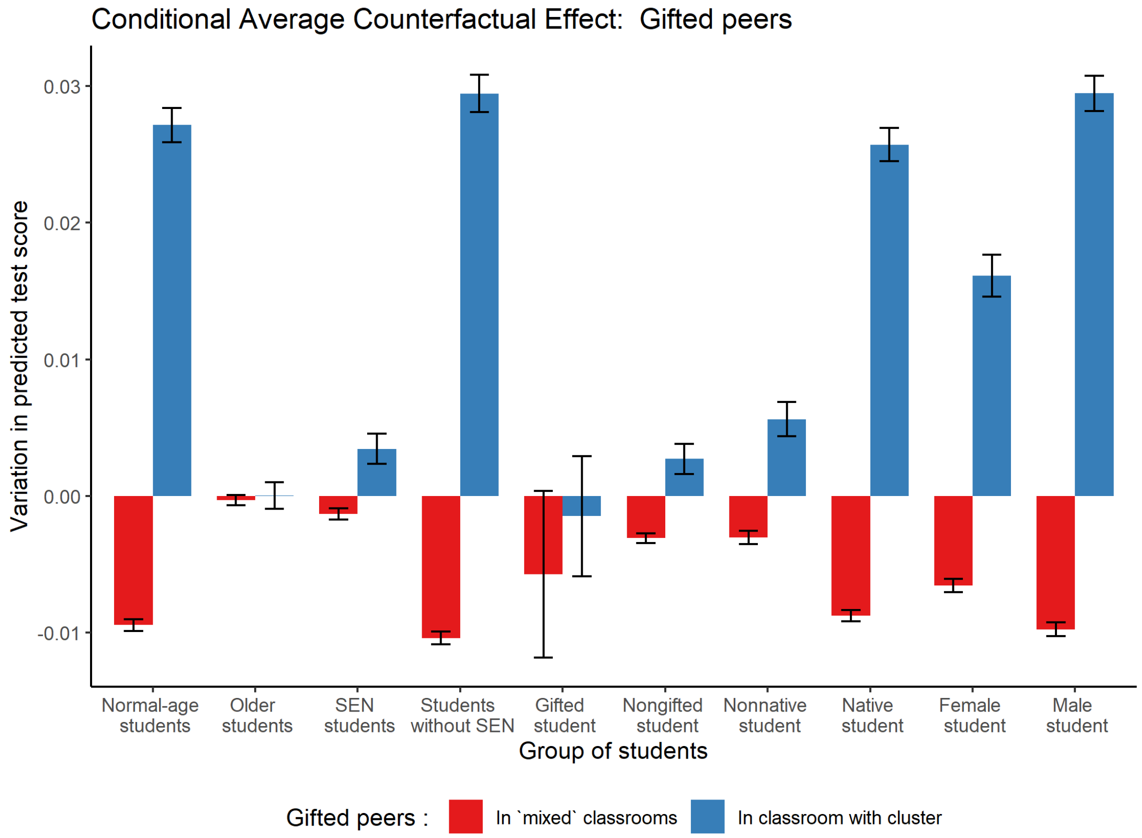
*Notes:* This figure displays the *conditional average counterfactual effects (CACE)* of classroom allocation manipulation for the types of nonnative and female. Each bar represent the difference in group test score averages under the different counterfactual regimes for each particular type of students. 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 4.15: CACE: clustering of nonnative and female students

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 4.13a, 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., Pregaldini, Backes-Gellner, and Eisenkopf, 2020; Eisenkopf et al., 2015). 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 4.17) generates positive peer effects 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



*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).

Figure 4.17: CACE: clustering of gifted students

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.

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.)

## 4.6 Conclusion

This study gives a more realistic understanding of peer effects among students, and integrates two limitations that were not sufficiently taken into account by the existing literature. First, peer effects have heterogeneous impacts on individuals with different characteristics. Second, there are *at least* as many peer effects as there are types of students. To account for these two elements in our analysis, we build on the assumption that "true" peer 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 peer effects influence academic performance the most, at learning about the way peer effects 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 peer 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. Our results also reveal that substantial heterogeneity hides between main effects, and that selected peer 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 non-native students, are downward sloping and not constant in the share of peers within classrooms. Finally, we simulate the marginal segregation of students into classroom clusters in order to empirically investigate deviations from full inclusive school setups. Marginally increasing segregation has different impacts depending on which students are marginally segregated. For instance, our simulations show that creating clusters of students with SEN, a population of students traditionally segregated, has a significant 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. Classroom allocations that "spread" students of all types across classrooms as much as possible, and doing so "distribute" the peers who generate positive spillovers, are found to be the allocations with highest aggregate school performance. Finally, while broadening our understanding of how to best serve students in inclusive school settings through meaningful and policy-relevant simulation exercises, this study invites for further research and more complete policy analysis regarding the influence of teachers and school resources.

# 5 | Conclusion: is inclusive schooling only a "rhetorical masterpiece"?

In his discussion of inclusive schooling, Haug (2017) argues that inclusive schooling was motivated solely on normative premises: the decision to implement inclusion was "value- and ideology-driven", a pure "rhetorical masterpiece". The arguments in favor of inclusion derived from the belief that a single school environment would better foster equal rights and equal opportunities. Inclusion was not thought as a policy whose efficiency can be measured and evaluated with empirical tools.[85] In that regard, one of the main challenges of the implementation of inclusive policies that Haug (2017) identifies in his theoretical discussion of inclusion is that inclusive policies have not been sufficiently and thoroughly empirically evaluated.

The three chapters of this thesis show that, indeed, inclusion is a policy that can be empirically evaluated. Moreover, these chapters incorporate the reality that inclusion is multi-faceted, and that it affects all students, not only students who would be otherwise segregated. Chapter 2 investigates special education programs for students with SEN in inclusive educational settings. It shows that inclusive programs are effective at generating academic success and labor market integration for students with SEN: returns to SpEd programs in inclusive settings (counseling and individual therapies) are positive for academic performance, and inclusion pays off in terms of academic performance, labor market participation and earnings in comparison to semi-segregation. Even though inclusion is better for (almost) all students with SEN, semi-segregation is almost as good as inclusion for SEN students who exhibit disruptive tendencies. Finally, simulations show that by mainstreaming most (if not all) students with SEN who would have been assigned to semi-segregation, we can reach higher academic and labor-market returns at lower costs.

Chapter 3 sheds light on the relevance of gifted students and their heterogeneous spillovers effects in the inclusive classroom. Our results show that, while male students benefit from the presence of gifted peers in all subjects regardless of gender, fe-

[85] In Haug (2017)'s own words: "I refer to inclusion as a masterpiece of rhetoric, easy to accept and difficult to be against or even criticize. This illustrates that inclusion is strongly value- and ideology-driven, in the same category as other similar concepts such as democracy and social justice."

male students benefit primarily from the presence of gifted female students. Moreover, exposure to gifted students is found to have powerful, lasting effects on career choices and post-compulsory education. Thus, gifted students are influential in fostering emulation and positively impacting the academic achievement and the career choices of their peers. They are therefore fundamental forces in the inclusive classroom production function which should not be ignored when assessing inclusive schooling as a policy, especially when considering whether gifted students should be segregated into more "elite" schools or pull-out programs.

Chapter 4 deploys new estimation techniques to more closely understand what happens in the mainstream classroom. The main insights of this chapter show that not all peer effects are equally important in the education production function. Moreover, we simulate the marginal segregation of students into clusters in one particular classroom in order to empirically investigate deviations from full inclusive school setups. Marginally increasing segregation has different impacts depending on which students are segregated. Classroom allocations that "spread" students of all types across classrooms and that "distribute" the peers that generate positive spillovers as much as possible are found to be the allocations with highest aggregate school performance. For instance, our simulations show that creating clusters of students with SEN, a population of students traditionally segregated, has a significant negative impact on the aggregated school performance.

Although many aspects of inclusive schooling are investigated in this thesis, some aspects would benefit from further research. In particular, the role of teachers and school resources in successfully implementing inclusive measures needs to be further investigated. Inclusion varies substantially in the way it is implemented, and empirical evidence on how teachers contribute to successful inclusive measures is lacking. Furthermore, promoters of inclusion argue that inclusion has beneficial effects not only on school performance and long-term human capital formation, but also on the well-being and altruism of all students. These claims need to be scrutinized empirically.

All in all, this thesis supports the empirical conclusion that inclusion works, and is beneficial on average and to most students in terms of academic performance and labor market outcomes. Shedding light on different angles of inclusive education, it adds empirical substance to this "rhetorical masterpiece". By building an empirical case that integrates different perspectives about inclusive education, the conclusions of each chapter should cumulatively convey, in Hume's own words, the "confidence" and "universal assent" necessary to claim that inclusion might become an "empirical masterpiece" as well.

# 6 | Bibliography

Alan, Sule, Seda Ertac, and Ipek Mumcu. 2018. "Gender stereotypes in the classroom and effects on achievement." *Review of Economics and Statistics* 100 (5):876–890.

Anelli, Massimo and Giovanni Peri. 2017. "The effects of high school peers' gender on college major, college performance and income." *Economic Journal* 129 (618):553–602.

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.

Antshel, Kevin, Stephen Faraone, Katharine Maglione, Alysa Doyle, Ronna Fried, Larry Seidman, and Joseph Biederman. 2008. "Temporal stability of ADHD in the high-IQ population: results from the MGH Longitudinal Family Studies of ADHD." *Journal of the American Academy of Child & Adolescent Psychiatry* 47 (7):817–825.

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, Julie Tibshirani, and Stefan Wager. 2019. "Generalized random forests." *Annals of Statistics* 47 (2):1148–1178.

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

———. 2021. "Policy learning with observational data." *Econometrica* 89 (1):133–161.

Avilova, Tatyana and Claudia Goldin. 2018. "What can UWE do for economics?" *AEA Papers & Proceedings* 108:186–190.

Avramidis, Elias and Brahm Norwich. 2002. "Teachers' attitudes towards integration/inclusion: a review of the literature." *European journal of special needs education* 17 (2):129–147.

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* .

Ballis, Briana and Katelyn Heath. forthcoming. "The long-run impacts of special education." *American Economic Journal: Economic Policy* 43.

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.

Bianco, Margarita, Bryn Harris, Dorothy Garrison-Wade, and Nancy Leech. 2011. "Gifted girls: Gender bias in gifted referrals." *Roeper Review* 33 (3):170–181.

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.

Blachman, Benita A, Christopher Schatschneider, Jack M Fletcher, Maria S Murray, Kristen A Munger, and Michael G Vaughn. 2014. "Intensive reading remediation in grade 2 or 3: Are there effects a decade later?" *Journal of educational psychology* 106 (1):46.

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.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of Machine Learning Research* 3 (Jan):993–1022.

Bloom, Harold, Carolyn Hill, A Black, and Mark Lipsey. 2006. "Effect sizes in education research: What they are, what they mean, and why they're important." *Institute of Education Sciences 2006 Research Conference, Washington, DC* .

Booij, Adam, Ferry Haan, and Erik Plug. 2016. "Enriching students pays off: Evidence from an individualized gifted and talented program in secondary education." Iza discussion paper no. 9757, IZA.

Booij, Adam, Edwin Leuven, and Hessel Oosterbeek. 2017. "Ability peer effects in university: evidence from a randomized experiment." *Review of Economic Studies* 84 (2):547–578.

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.

Brown, Thomas, Philipp Reichel, and Donald Quinlan. 2009. "Executive function impairments in high IQ adults with ADHD." *Journal of Attention Disorders* 13 (2):161–167.

Buckles, Kasey. 2019. "Fixing the leaky pipeline: strategies for making economics work for women at every stage." *Journal of Economic Perspectives* 33 (1):43–60.

Bui, Sa, Steven Craig, and Scott Imberman. 2014. "Is gifted education a bright idea? Assessing the impact of gifted and talented programs on students." *American Economic Journal: Economic Policy* 6 (3):30–62.

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

Buser, Thomas, Noemi Peter, and Stefan Wolter. 2017. "Gender, willingness to compete and career choices along the whole ability distribution." Iza discussion paper no. 10976, IZA.

Cappelen, Alexander, John List, Anya Samek, and Bertil Tungodden. 2020. "The effect of early-childhood education on social preferences." *Journal of Political Economy* 128 (7):2739–2758.

Card, David and Laura Giuliano. 2014. "Does gifted education work? For which students?" Nber working paper no. 20453, National Bureau of Economic Research.

———. 2016. "Universal screening increases the representation of low-income and minority students in gifted education." *Proceedings of the National Academy of Sciences* 113 (48):13678–13683.

Card, David and Abigail Payne. 2017. "High school choices and the gender gap in STEM." Nber working paper no. 23769, National Bureau of Economic Research.

Carlana, Michela. 2019. "Implicit stereotypes: Evidence from teachers' gender bias." *Quarterly Journal of Economics* 134 (3):1163–1224.

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

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.

Carrell, Scott, Marianne Page, and James West. 2010. "Sex and science: how professor gender perpetuates the gender gap." *Quarterly Journal of Economics* 125 (3):1101–1144.

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

Case, Anne, Darren Lubotsky, and Christina Paxson. 2002. "Economic Status and Health in Childhood: The Origins of the Gradient." *American Economic Review* 92 (5):1308–1334.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *Econometrics Journal* 21:C1–C68.

Chernozhukov, Victor, Iván Fernández-Val, and Ye Luo. 2018. "The sorted effects method: discovering heterogeneous effects beyond their averages." *Econometrica* 86 (6):1911–1938.

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

Cho, Rosa Minhyo. 2012. "Are there peer effects associated with having English language learner (ELL) classmates? Evidence from the Early Childhood Longitudinal Study Kindergarten Cohort (ECLS-K)." *Economics of Education Review* 31 (5):629–643.

Cole, Cassandra M, Nancy Waldron, and Massoumeh Majd. 2004. "Academic progress of students across inclusive and traditional settings." *Mental retardation* 42 (2):136–144.

Cools, Angela, Raquel Fernández, and Eleonora Patacchini. 2019. "Girls, boys, and high achievers." Nber working paper no. 25763, National Bureau of Economic Research.

Crump, Richard, Joseph Hotz, Guido Imbens, and Oscar Mitnik. 2009. "Dealing with limited overlap in estimation of average treatment effects." *Biometrika* 96 (1):187–199.

Cullen, Julie Berry. 2003. "The impact of fiscal incentives on student disability rates." *Journal of Public Economics* 87 (7-8):1557–1589.

Cunha, Flavio and James Heckman. 2007. "The technology of skill formation." *American Economic Review* 97 (2):31–47.

Currie, Janet and Mark Stabile. 2003. "Socioeconomic Status and Child Health: Why Is the Relationship Stronger for Older Children?" *American Economic Review* 93 (5):1813–1823.

Daniel, Larry G and Debra A King. 1997. "Impact of inclusion education on academic achievement, student behavior and self-esteem, and parental attitudes." *The Journal of Educational Research* 91 (2):67–80.

Davis, Jonathan M.V. and Sara B. Heller. 2017. "Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs." *American Economic Review* 107 (5):546–50.

De Boer, Anke, Sip Jan Pijl, and Alexander Minnaert. 2011. "Regular primary schoolteachers' attitudes towards inclusive education: A review of the literature." *International journal of inclusive education* 15 (3):331–353.

De Bruin, Kate. 2019. "The impact of inclusive education reforms on students with disability: an international comparison." *International Journal of Inclusive Education* 23 (7-8):811–826.

Deary, Ian, Steve Strand, Pauline Smith, and Cres Fernandes. 2007. "Intelligence and educational achievement." *Intelligence* 35 (1):13–21.

Dee, Thomas. 2007. "Teachers and the gender gaps in student achievement." *Journal of Human Resources* 42 (3):528–554.

Dempsey, Ian, Megan Valentine, and Kim Colyvas. 2016. "The Effects of Special Education Support on Young Australian School Students." *International Journal of Disability, Development and Education* 63 (3):271–292.

Der, Geoff, David Batty, and Ian Deary. 2009. "The association between IQ in adolescence and a range of health outcomes at 40 in the 1979 US National Longitudinal Study of Youth." *Intelligence* 37 (6):573–580.

Diette, Timothy M and Ruth Uwaifo Oyelere. 2014. "Gender and race heterogeneity: The impact of students with limited english on native students' performance." *American Economic Review* 104 (5):412–17.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer effects, teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya." *American Economic Review* 101 (5):1739–1774.

Duncan, Greg J. and Katherine Magnuson. 2013. "Investing in Preschool Programs." *Journal of Economic Perspectives* 27 (2):109–32.

Duncombe, William and John Yinger. 2005. "How much more does a disadvantaged student cost?" *Economics of Education Review* 24 (5):513– 532.

D'Amour, Alexander, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. 2021. "Overlap in observational studies with high-dimensional covariates." *Journal of Econometrics* 221 (2):644–654.

EASIE. 2014. "2014 Dataset Cross-Country Report." Tech. rep., European Agency Statistics on Inclusive Education.

———. 2018. "2018 Dataset Cross-Country Report." Tech. rep., European Agency Statistics on Inclusive Education.

Eckhart, Michael, Urs Haeberlin, Sahli Lozano Caroline, and Philippe Blanc. 2011. "Langzeitwirkungen der schulischen Integration. Eine empirische Studie zur Bedeutung von Integrationserfahrungen in der Schulzeit für die soziale und berufliche Situation im jungen Erwachsenenalter." *Bern, Stuttgart, Wien: Haupt* .

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2018. "How to Make Causal Inferences Using Texts." *ArXiv Working Paper* (1802.02163).

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.

Elder, Todd, David Figlio, Scott Imberman, and Claudia Persico. 2021. "School Segregation and Racial Gaps in Special Education Identification." *Journal of Labor Economics* Forthcoming.

Elder, Todd E. 2010. "The importance of relative standards in ADHD diagnoses: Evidence based on exact birth dates." *Journal of Health Economics* 29 (5):641– 656.

Ellison, Glenn and Ashley Swanson. 2010. "The gender gap in secondary school mathematics at high achievement levels: evidence from the American Mathematics Competitions." *Journal of Economic Perspectives* 24 (2):109–128.

Eugster, Beatrix and Raphaël Parchet. 2019. "Culture and taxes." *Journal of Political Economy* 127 (1):296–337.

Fan, Qingliang, Yu-Chin Hsu, Robert P Lieli, and Yichong Zhang. 2020. "Estimation of conditional average treatment effects with high-dimensional data." *Journal of Business & Economic Statistics* :1–15.

Farrell, Max H. 2015. "Robust inference on average treatment effects with possibly more covariates than observations." *Journal of Econometrics* 189 (1):1– 23.

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

Firpo, Sergio, Nicole Fortin, and Thomas Lemieux. 2009. "Unconditional quantile regressions." *Econometrica* 77 (3):953–973.

Fletcher, Jason and Barbara Wolfe. 2008. "Child mental health and human capital accumulation: The case of ADHD revisited." *Journal of Health Economics* 27 (3):794–800.

Fletcher, Jason M. 2009. “The effects of inclusion on classmates of students with special needs: The case of serious emotional problems.” *Education Finance and Policy* 4 (3):278–299.

Fletcher, Jason M. 2014. “The effects of childhood ADHD on adult labor market outcomes.” *Health Economics* 23 (2):159–181.

Freeman, Stephanny FN and Marvin C Alkin. 2000. “Academic and social attainments of children with mental retardation in general education and special education settings.” *Remedial and Special Education* 21 (1):3–26.

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

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. “Text as data.” *Journal of Economic Literature* 57 (3):535–74.

Glynn, Adam N. and Kevin M. Quinn. 2010. “An Introduction to the Augmented Inverse Propensity Weighted Estimator.” *Political Analysis* 18 (1):36–56.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. “Performance in competitive environments: Gender differences.” *Quarterly Journal of Economics* 118 (3):1049–1074.

Gomez, Rapson, Vasileios Stavropoulos, Alasdair Vance, and Mark Griffiths. 2019. “Gifted children with ADHD: how are they different from non-gifted children with ADHD?” *International Journal of Mental Health and Addiction* :1–15.

Gottfried, Allen, Adele Eskeles Gottfried, Kay Bathurst, and Diana Wright Guerin. 1994. *Gifted IQ: Early Developmental Aspects – The Fullerton Longitudinal Study*. Springer Science & Business Media.

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.

Greminger, Eva, Rupert Tarnutzer, and Martin Venetz. 2005. “Die Tragfähigkeit der Regelschule stärken.” *Schweizerische Zeitschrift für Heilpädagogik, 7* 8 (5):49–52.

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.

Häfeli, Kurt and Peter Walther-Müller. 2005. "Das Wachstum des sonderpädagogischen Angebots im interkantonalen Vergleich." *Schweizerische Zeitschrift für Heilpädagogik* (7-8).

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.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. 2002. "Inferring Program Effects for Special Populations: Does Special Education Raise Achievement for Students with Disabilities?" *The Review of Economics and Statistics* 84 (4):584–599.

Harrison, Judith R., Nora Bunford, Steven W. Evans, and Julie Sarno Owens. 2013. "Educational Accommodations for Students With Behavioral Challenges: A Systematic Review of the Literature." *Review of Educational Research* 83 (4):551–597.

Haug, Peder. 2017. "Understanding inclusive education: ideals and reality." *Scandinavian Journal of Disability Research* 19 (3):206–217.

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103 (6):2052–86.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz. 2010. "The rate of return to the HighScope Perry Preschool Program." *Journal of Public Economics* 94 (1):114–128.

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.

Hoxby, Caroline. 2000. "Peer effects in the classroom: learning from gender and race variation." Nber working paper no. 7867, National Bureau of Economic Research.

Hyde, Janet and Janet Mertz. 2009. "Gender, culture, and mathematics performance." *Proceedings of the National Academy of Sciences* 106 (22):8801–8807.

Imbens, Guido W. 2000. "The role of the propensity score in estimating dose-response functions." *Biometrika* 87 (3):706–710.

Imbens, Guido W and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Iriberri, Nagore and Pedro Rey-Biel. 2019. "Competitive pressure widens the gender gap in performance: Evidence from a two-stage competition in mathematics." *Economic Journal* 129 (620):1863–1893.

Isphording, Ingo and Ulf Zölitz. 2020. "The value of a peer." Department of economics working paper no. 342, University of Zurich.

Jackson, Kirabo, Rucker Johnson, and Claudia Persico. 2016. "The effects of school spending on educational and economic outcomes: Evidence from school finance reforms." *Quarterly Journal of Economics* 131 (1):157–218.

Judge, Sharon and Silvana M. R. Watson. 2011. "Longitudinal Outcomes for Mathematics Achievement for Students with Learning Disabilities." *The Journal of Educational Research* 104 (3):147–157.

Karpinski, Ruth, Audrey Kinase Kolb, Nicole Tetreault, and Thomas Borowski. 2018. "High intelligence: A risk factor for psychological and physiological overexcitabilities." *Intelligence* 66:8–23.

Keith, Katherine A, David Jensen, and Brendan O'Connor. 2020. "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates." *arXiv preprint arXiv:2005.00649* .

Kennedy, Edward H. 2020. "Optimal doubly robust estimation of heterogeneous causal effects." *arXiv preprint arXiv:2004.14497* .

Keslair, Francois, Eric Maurin, and Sandra McNally. 2012. "Every child matters? An evaluation of "Special Educational Needs? programmes in England." *Economics of Education Review* 31 (6):932– 948.

Kirjavainen, Tanja, Jonna Pulkkinen, and Markku Jahnukainen. 2016. "Special education students in transition to further education: A four-year register-based follow-up study in Finland." *Learning and Individual Differences* 45:33– 42.

Kirkeboen, Lars J, Edwin Leuven, and Magne Mogstad. 2016. "Field of study, earnings, and self-selection." *The Quarterly Journal of Economics* 131 (3):1057–1111.

Kitagawa, Toru and Aleksey Tetenov. 2018. "Who should be treated? empirical welfare maximization methods for treatment choice." *Econometrica* 86 (2):591–616.

Knaus, Michael. 2021. "Double Machine Learning based Program Evaluation under Unconfoundedness." *Working Paper* .

Knaus, Michael, Michael Lechner, and Anthony Strittmatter. 2020. "Heterogeneous employment effects of job search programmes: A machine learning approach." *Journal of Human Resources* :0718–9615R1.

Kohli, Nidhi, Amanda L. Sullivan, Shanna Sadeh, and Cengiz Zopluoglu. 2015. "Longitudinal mathematics development of students with learning disabilities and students without disabilities: A comparison of linear, quadratic, and piecewise linear mixed effects models." *Journal of School Psychology* 53 (2):105– 120.

Kvande, Marianne Nilsen, Oda Bjørklund, Stian Lydersen, Jay Belsky, and Lars Wichstrøm. 2018. "Effects of special education on academic achievement and task motivation: a propensity-score and fixed-effects approach." *European Journal of Special Needs Education* 0 (0):1–15.

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. 2005. "Targeted Remedial Education for Underperforming Teenagers: Costs and Benefits." *Journal of Labor Economics* 23 (4):839–874.

———. 2011a. "Mechanisms and impacts of gender peer effects at school." *American Economic Journal: Applied Economics* 3 (2):1–33.

———. 2011b. "Mechanisms and impacts of gender peer effects at school." *American Economic Journal: Applied Economics* 3 (2):1–33.

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

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

Lechner, Michael. 2001. "Identification and estimation of causal effects of multiple treatments under the conditional independence assumption." In *Econometric Evaluation of Labour Market Policies*. Physica-Verlag HD, 43–58.

———. 2019. "Modified causal forests for estimating heterogeneous causal effects." *arXiv preprint arXiv:1812.09487* .

Lekhal, Ratib. 2018. "Does special education predict students' math and language skills?" *European Journal of Special Needs Education* 33 (4):525–540.

Li, Fan and Fan Li. 2019. "Propensity score weighting for causal inference with multiple treatments." *The Annals of Applied Statistics* 13 (4):2389–2415.

Li, Fan, Kari Lock Morgan, and Alan M Zaslavsky. 2018. "Balancing covariates via propensity score weighting." *Journal of the American Statistical Association* 113 (521):390–400.

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

Lovett, Maureen W, Jan C Frijters, Maryanne Wolf, Karen A Steinbach, Rose A Sevcik, and Robin D Morris. 2017. "Early intervention for children at risk for reading disabilities: The impact of grade at intervention and individual differences on intervention outcomes." *Journal of Educational Psychology* 109 (7):889.

Lynn, Richard and Gerhard Meisenberg. 2010. "National IQs calculated and validated for 108 nations." *Intelligence* 38 (4):353–360.

Lynn, Richard and Tatu Vanhanen. 2012. "National IQs: a review of their educational, cognitive, economic, political, demographic, sociological, epidemiological, geographic and climatic correlates." *Intelligence* 40 (2):226–234.

Maestas, Nicole, Kathleen J Mullen, and Alexander Strand. 2013. "Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt." *American Economic Review* 103 (5):1797–1829.

Mahone, Mark, Kathleen Hagelthorn, Laurie Cutting, Linda Schuerholz, Shelley Pelletier, Christine Rawlins, Harvey Singer, and Martha Denckla. 2002. "Effects of IQ on executive function measures in children with ADHD." *Child Neuropsychology* 8 (1):52–65.

Make, Matthew and Jonathan Plucker. 2018. "Creativity." In *Handbook of Giftedness in Children: Psychoeducational Theory, Research, and Best Practices*, edited by Steven Pfeiffer. Springer International Publishing, 247–270".

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

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

———. 2004. "Statistical treatment rules for heterogeneous populations." *Econometrica* 72 (4):1221–1246.

Mansour, Hani, Daniel Rees, Bryson Rintala, and Nathan Wozny. 2018. "The effects of professor gender on the post-graduation outcomes of female students." Iza discussion paper no. 11820, IZA.

Mayer, John, Donna Perkins, David Caruso, and Peter Salovey. 2001. "Emotional intelligence and giftedness." *Roeper Review* 23 (3):131–137.

McDermott, Paul, Marley Watkins, and Anna Rhoad. 2014. "Whose IQ is it? Assessor bias variance in high-stakes psychological assessment." *Psychological Assessment* 26 (1):207–214.

McGee, Andrew. 2011. "Skills, standards, and disabilities: How youth with learning disabilities fare in high school and beyond." *Economics of Education Review* 30 (1):109–129.

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

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. “Distributed representations of words and phrases and their compositionality.” In *Advances in neural information processing systems*. 3111–3119.

Montolio, Daniel and Pere Taberner. 2018. “Gender differences under test pressure and their impact on academic performance: a quasi-experimental design.” Ieb working paper 2018/21, IEB.

Morgan, Paul L., George Farkas, and Qiong Wu. 2009. “Five-Year Growth Trajectories of Kindergarten Children With Learning Difficulties in Mathematics.” *Journal of Learning Disabilities* 42 (4):306–321.

Morgan, Paul L., Michelle L. Frisco, George Farkas, and Jacob Hibel. 2010. “A Propensity Score Matching Analysis of the Effects of Special Education Services.” *The Journal of Special Education* 43 (4):236–254.

Morgenroth, Thekla, Michelle Ryan, and Kim Peters. 2015. “The motivational theory of role modeling: How role models influence role aspirants’ goals.” *Review of General Psychology* 19 (4):465–483.

Morin, Louis-Philippe. 2015. “Do men and women respond differently to competition? Evidence from a major education reform.” *Journal of Labor Economics* 33 (2):443–491.

Mouganie, Pierre and Yaojing Wang. 2020. “High performing peers and female STEM choices in school.” *Journal of Labor Economics* 38 (3):805–841.

Mozer, Reagan, Luke Miratrix, Aaron Russell Kaufman, and L. Jason Anastasopoulos. 2020. “Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality.” *Political Analysis* 28 (4):445–468.

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

Neisser, Ulric, Gwyneth Boodoo, Thomas Bouchard, Wade Boykin, Nathan Brody, Stephen Ceci, Diane Halpern, John Loehlin, Robert Perloff, Robert Sternberg, and Susana Urbina. 1996. “Intelligence: knowns and unknowns.” *American Psychologist* 51 (2):77–101.

Newman, Tina. 2008. “Assessment of giftedness in school-age children using measures of intelligence or cognitive abilities.” In *Handbook of Giftedness in Children: Psychoeducational Theory, Research, and Best Practices*, edited by Steven Pfeiffer. Springer International Publishing, 161–176".

Niederle, Muriel and Lise Vesterlund. 2007. “Do women shy away from competition? Do men compete too much?” *Quarterly Journal of Economics* 122 (3):1067–1101.

———. 2010. “Explaining the gender gap in math test scores: The role of competition.” *Journal of Economic Perspectives* 24 (2):129–144.

———. 2011. “Gender and competition.” *Annual Review of Economics* 3 (1):601–630.

Nosek, Brian, Frederick Smyth, Natarajan Sriram, Nicole Lindner, Thierry Devos, Alfonso Ayala, Yoav Bar-Anan, Robin Bergh, Huajian Cai, Karen Gonsalkorale, Selin Kesebir, Norbert Maliszewski, Félix Neto, Eero Olli, Jaihyun Park, Konrad Schnabel, Kimihiro Shiomura, Bogdan Tudor Tulbure, Reinout Wiers, Mónika Somogyi, Nazar Akrami, Bo Ekehammar, Michelangelo Vianello, Mahzarin Banaji, and Anthony Greenwald. 2009. “National differences in gender-science stereotypes predict national sex differences in science and math achievement.” *Proceedings of the National Academy of Sciences* 106 (26):10593–10597.

OECD. 2017. *Education at a Glance 2017: OECD Indicators*. Organisation for Economic Cooperation and Development, Paris (France).

Peetsma, Thea, Margaretha Vergeer, Jaap Roeleveld, and Sjoerd Karsten. 2001. “Inclusion in education: Comparing pupils’ development in special and regular education.” *Educational Review* 53 (2):125–135.

Peters, Scott and Michael Stuart Matthews. 2016. “Gifted education research from the economists’ perspective: What have we learned?” *Journal of Advanced Academics* 27 (2):150–161.

Petersen, Jennifer. 2013. “Gender differences in identification of gifted youth and in gifted program participation: A meta-analysis.” *Contemporary Educational Psychology* 38 (4):342–348.

Peyre, Hugo, Franck Ramus, Maria Melchior, Anne Forhan, Barbara Heude, and Nicolas Gauvrit. 2016. “Emotional, behavioral and social difficulties among high-IQ children during the preschool period: Results of the EDEN mother-child cohort.” *Personality and Individual Differences* 94:366–371.

Pope, Devin and Justin Sydnor. 2010. “Geographic variation in the gender differences in test scores.” *Journal of Economic Perspectives* 24 (2):95–108.

Porter, Catherine and Danila Serra. 2019. “Gender differences in the choice of major: The importance of female role models.” *American Economic Journal: Applied Economics* .

Preckel, Franzis, Thomas Goetz, Reinhard Pekrun, and Michael Kleine. 2008. “Gender differences in gifted and average-ability students: Comparing girls’ and boys’ achievement, self-concept, interest, and motivation in mathematics.” *Gifted Child Quarterly* 52 (2):146–159.

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.

Rangvid, Beatrice Schindler. 2019. “Returning special education students to regular classrooms: Externalities on peers’ reading scores.” *Economics of Education Review* 68:13–22.

Roberts, Margaret E, Brandon M Stewart, and Edoardo M Airoldi. 2016. “A model of text for experimentation in the social sciences.” *Journal of the American Statistical Association* 111 (515):988–1003.

Roberts, Margaret E, Brandon M Stewart, and Richard A Nielsen. 2020. “Adjusting for confounding with text matching.” *American Journal of Political Science* .

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. “Estimation of Regression Coefficients When Some Regressors are not Always Observed.” *Journal of the American Statistical Association* 89 (427):846–866.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 1995. “Analysis of semiparametric regression models for repeated outcomes in the presence of missing data.” *Journal of the american statistical association* 90 (429):106–121.

Rosenbaum, Paul R. and Donald B. Rubin. 1985. “Constructing a control group using multivariate matched sampling methods that incorporate the propensity score.” *The American Statistician* 39 (1):33–38.

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* .

Saygin, Perihan. 2020. “Gender bias in standardized tests: evidence from a centralized college admissions system.” *Empirical Economics* 59:1037–1065.

Schwab, Susanne. 2020. “Inclusive and special education in Europe.” In *Oxford Research Encyclopedia of Education*.

Schwartz, Amy Ellen, Bryant Gregory Hopkins, and Leanna Stiefel. 2021. “The Effects of Special Education on the Academic Performance of Students with Learning Disabilities.” *Journal of Policy Analysis and Management* 40 (2):480–520.

Scruggs, Thomas E., Margo A. Mastropieri, Sheri Berkeley, and Janet E. Graetz. 2010. “Do Special Education Interventions Improve Learning of Secondary Content? A Meta-Analysis.” *Remedial and Special Education* 31 (6):437–449.

Semenova, Vira and Victor Chernozhukov. 2021. “Debiased machine learning of conditional average treatment effects and other causal functions.” *The Econometrics Journal* 24 (2):264–289.

Sermier-Dessemontet, Rachel, Valérie Benoit, and Gérard Bless. 2011. "Schulische Integration von Kindern mit einer geistigen Behinderung. Untersuchung der Entwicklung der Schulleistungen und der adaptiven Fähigkeiten, der Wirkung auf die Lernentwicklung der Mitschüler sowie der Lehrereinstellungen zur Integration." *Empirische Sonderpädagogik* 3 (4):291–307.

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.

Silverman, Linda Kreger. 2018. "Assessment of giftedness." In *Handbook of Giftedness in Children: Psychoeducational Theory, Research, and Best Practices*, edited by Steven Pfeiffer. Springer International Publishing, 183–207".

Smith, James P. 2009. "The Impact of Childhood Health on Adult Labor Market Outcomes." *The Review of Economics and Statistics* 91 (3):478–489.

Sternberg, Robert, Linda Jarvin, and Elena Grigorenko. 2010. *Explorations in Giftedness*. Cambridge University Press.

Stürmer, Til, Kenneth J Rothman, Jerry Avorn, and Robert J Glynn. 2010. "Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study." *American Journal of Epidemiology* 172 (7):843–854.

Sullivan, Amanda L. and Samuel Field. 2013. "Do preschool special education services make a difference in kindergarten reading and mathematics skills?: A propensity score weighting analysis." *Journal of School Psychology* 51 (2):243– 260.

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

Van der Laan, Mark J, Eric C Polley, and Alan E Hubbard. 2007. "Super learner." *Statistical applications in genetics and molecular biology* 6 (1).

Vardardottir, Arna. 2015. "The impact of classroom peers in a streaming system." *Economics of Education Review* 49:110–128.

Weld, Galen, Peter West, Maria Glenski, David Arbour, Ryan Rossi, and Tim Althoff. 2020. "Adjusting for Confounders with Text: Challenges and an Empirical Evaluation Framework for Causal Inference." *arXiv preprint arXiv:2009.09961* .

Whitmore, Diane. 2005. "Resource and peer impacts on girls' academic achievement: evidence from a randomized experiment." *AEA Papers & Proceedings* 95 (2):199–203.

Wolter, Stefan C. and Miriam Kull. 2006. "Bildungsbericht 2006." Tech. rep. URL `http://www.skbf-csre.ch/bildungsbericht/bildungsbericht/`.

———. 2014. “Bildungsbericht 2014.” Tech. rep. URL `http://www.skbf-csre.ch/bildungsbericht/bildungsbericht/`.

Xu, Guifeng, Lane Strathearn, Buyun Liu, Binrang Yang, and Wei Bao. 2018. “Twenty-year trends in diagnosed attention-deficit/hyperactivity disorder among US children and adolescents, 1997-2016.” *JAMA network open* 1 (4):e181471–e181471.

Yoshida, Kazuki, Daniel H Solomon, Sebastien Haneuse, Seoyoung C Kim, Elisabetta Patorno, Sara K Tedeschi, Houchen Lyu, Jessica M Franklin, Til Stürmer, Sonia Hernández-Díaz et al. 2019. “Multinomial extension of propensity score trimming methods: a simulation study.” *American Journal of Epidemiology* 188 (3):609–616.

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.

Zeidner, Moshe, Inbal Shani-Zinovich, Gerald Matthews, and Richard Roberts. 2005. “Assessing emotional intelligence in gifted and non-gifted high school students: outcomes depend on the measure.” *Intelligence* 33 (4):369–391.

Zhou, Zhengyuan, Susan Athey, and Stefan Wager. 2018. “Offline multi-action policy learning: Generalization and optimization.” *arXiv preprint arXiv:1810.04778* .

Zimmert, Michael and Michael Lechner. 2019. “Nonparametric estimation of causal heterogeneity under high-dimensional confounding.” *arXiv preprint arXiv:1908.08779* .

Zölitz, Ulf and Jan Feld. 2021. “The effect of peer gender on major choice in business school.” *Management Science* 67 (11):6963–6979.

# A | Appendix: Chapter 2

## A.1 Appendix: Using text to adjust for confounding

The purpose of using computational text analysis methods and Natural Language Processing (NLP) in this paper is for confounding adjustment in the estimation of returns to SE programs. It secondarily serves as an interesting pretreatment variable to explore treatment heterogeneity.

Extracting information from raw text is difficult for two main reasons: first, text is unstructured and high-dimensional, and, second, it includes latent features. To account for these two difficulties, text must be represented by an unknown $g()$ function that must be discovered in order to make text comparable and interpretable, as well as compress its dimensionality to a lower dimensional space. The main trade-off in discovering $g()$ is to compress the high-dimensionality of text without losing its substantial meaning: marginally extracting more information from the text occurs at the cost of increasing the dimensions of the covariate space, which leads to support and computational problems. Traditionally, $g()$ is discovered by human coders who extract relevant dimensions of the text into a low-dimensional space (e.g., a series of indicators). Recent machine-learning methods discover $g()$ in an unsupervised manner (e.g., Blei, Ng, and Jordan (2003)).

This appendix presents in greater details the way I tackle these two difficulties and how I prepare the written psychological records for the estimation of treatment effects. It provides information on how I preprocessed the text, as well as how I implement the different methods. Summary statistics about the distribution of text statistics across treatment states are also provided.

### A.1.1 Text preprocessing

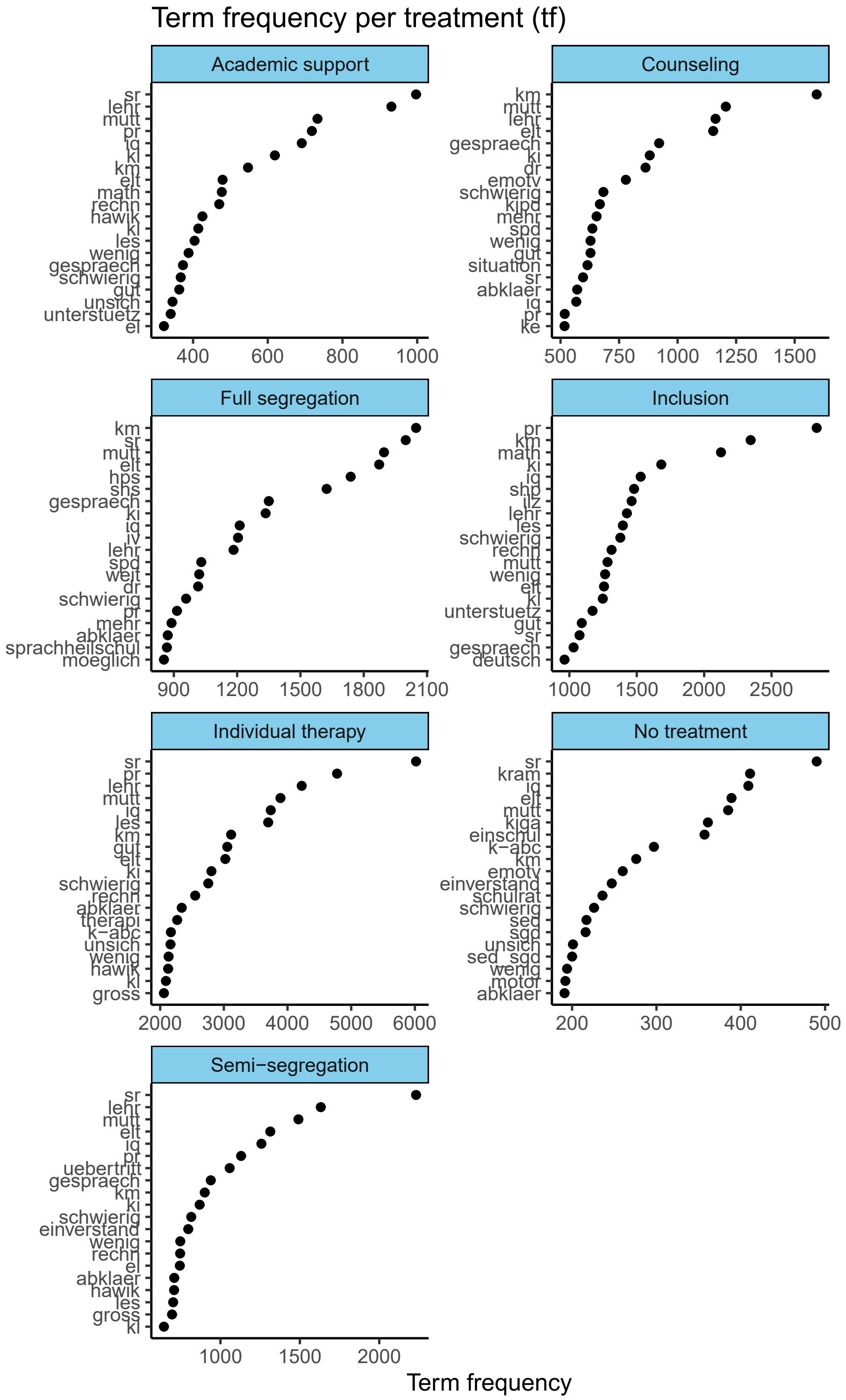
Most text analysis methods implemented in this study require the text to be preprocessed — with the exception of word embeddings. For preprocessing, I strip the text from words which have low information value (so-called "stopwords", numbers, and punctuation signs). Subsequently, I reduce the text to single words ("tokens"), and lemmatize them.(for more details, see Manning, Raghavan, and Schütze, 2008). The lemmatization ensures that all tokens are reduced to their stems/roots without inflectional endings. Finally, I represent the text into types, i.e. unique expressions of tokens made out of onegrams (single tokens) and bigrams (two co-occuring tokens). For instance, if the number of tokens in the following document "I ate an apple apple" is 5, the number of one-gram types is 4, as the token "apple" is repeated. After preprocessing, the psychological records per student contain on average 241 tokens and 137 types.

### A.1.2 Text representations

**Document term matrix (DTM)** The simplest way of representing a text is to reduce the text to its components ("tokens") and, for each token, indicate its frequency of appearance within each document in a "Document term matrix" (DTM). Thus, two documents are identical (and comparable) if they use the same tokens with a similar frequency. The limitations of DTMs is that frequency matrices do not account for the context in which tokens appear. Moreover, the dimensionality of DTMs quickly explodes with the number of documents, making comparisons across documents difficult. This requires handling huge sparse matrices, which can be computationally difficult.

To reduce the dimension of the DTM, I implement two methods: scaling and bounding. For scaling, I weight the DTM either by term frequency *tf*, which simply weights terms according to their number of appearances in a document, or by term-frequency-inverse document frequency (*tf-idf*), which increases with the term frequency of a word in a document and decreases with the number of documents in the corpus that contains the word (thus highlighting words that carry a lot of information about the document). Basically, very often used "stop words" have a very low *tf-idf* score, and words that are highly specific to a document have high *tf-idf* scores.

To select terms that are general enough, I bound the number of terms by selecting words that appear at least 350 times (min. term frequence) and in at least 150 documents (min doc. frequency) for *tf*. For *tf-idf*, I select very specific tokens, i.e. *tf-idf* score bounded at the 99.9th percentile of all *tf-idf* scores. I provide a third measure with a mixture of *tf* and *tf-idf* scores: I first select the most frequent tokens, and then weight them by *tf-idf*, which ensures that only frequent words with high significance are se-



*Notes:* This figure represents the 20 most frequent terms (tokens) per treatment assignment. Text is preprocessed via lemmatization. *Source: SPS.*

Figure A.1: Most common tokens per treatment assignment

lected. As an illustration, Figure A.1 and Figure A.2 display the 20 most frequent terms per treatment assignment.

The criterion for the choice of scaling scheme and tuning parameters in the DTM representation depends on the empirical problem at hand: I do not only want tokens that are very predictive of treatment and outcome, but also tokens that are common enough to serve the purpose of comparing documents. Thus, a very high *tf-idf* may suit the purpose of prediction very well; however, scores in the middle range might serve the purpose of comparison better.

**Structural Topic Modeling (STM) and Topical Inverse Regression Matching (TIRM)** To mitigate the high-dimensionality and lack of word context in DTM, representations that discover latent features of the text are interesting ways of representing high-dimensional text. I implement Structural Topic Models (STM) as proposed by Roberts, Stewart, and Airoldi (2016) and Topical Inverse Regression Matching (TIRM) Roberts, Stewart, and Nielsen (2020), which are a variant of the Latent Dirichlet Allocation topic model (LDA, Blei, Ng, and Jordan (2003)). Succinctly, LDA and STM first sample a topic from the distribution of tokens in a given document, and then sample each observed token from the distribution of words given each topic. Unlike LDA, STM allows for covariates to affect the proportion of a document attributed to a topic ("topical prevalence") and the distribution of tokens within a topic ("topical content"). STM assumes that covariates influence the way tokens are distributed: it includes covariates in the prior distribution of topics per documents (logistic normal distribution) and the distribution of tokens per topic (multinomial logistic regression). In STM, the number of topics is chosen *ad-hoc*, and topics are not directly interpretable.

The advantage of STM is that it reduces the text dimensionality into a finite number of topics, and offers a good comparison of documents, as documents that cover the same topics at the same rates are represented as similar documents. In this application, I estimate STM and define the number of topics $k$ to either 10 or 80. I then use the vector of $k$ topic proportions ($k \times N$) directly in the propensity score. Other variants are possible, such as taking the $k - x$ most important topics or the topics that explain topical content the most (see, for instance, Mozer et al., 2020).

Topical Inverse Regression Matching (TIRM) is an interesting method that allows to directly model the treatment assignment process from text. It builds on STM by estimating an additional reduction, i.e. a document-level propensity score based on

| Topic | Highest probability | Most frequent and exclusive |
|---|---|---|
| 1 | sr, mutt, km, schulrat, les, pr, kiga, vgl, lehr, pr | sr_lekt, vb, jug, u'ergebnis, iq_sed, gemeinsam_auswertungsgespraech, trog-d, vgl_notiz, proz, sht |
| 2 | iq, elt, ki, besprech, rechn, hawik, einschul, notiz, ke, kl | untersuch_hawik, herrn, inform, besprech_-u'ergebnis, auditiv_merkfaeh, psychodiagnost_gemeinsam, wwt, notiz_vgl, sek, leistungs-_lernverhalt |
| 3 | lekt, gespraech, kv, elt, k-abc, iq, shs, mutt, mutt, elt | untersuch_k-abc, bad_sond, beob, leg.-th, dyskalkulie-therapi, wld_agd, einschulungsjahr, vgl, ke, pl |
| 4 | hawik, dr, math, abklaer, sed, sv, motor, ki, schwierig, iq | lekt_vorlaeuf, beistand, z.h, abklaer_wunsch, forst, agd_vg, hpd, antragsschreib, li, lernverhalt_aktuell |
| 5 | untersuch, spd, jedoch, mutt, sgd, vg, auffaell, vgl_notiz, mehr, gut | testsitz, km, semesterbericht, macht_mueh, agd, shs, notiz, befind, lernverhalt |
| 6 | lehr, kjpd, sp, lehrerin, sed_sgd, wld, abklaer, lehr, bess, cpm | rav_pr, moegl, dc-ther, les_langsam, kontextklaer, hs_ds, vg_mutt, familia, cpm_rav |
| 7 | schulleist, weit, kl, vorgespraech, unsich, agd, kg, abkl, wenig, kram | kkd, langhald, zz, wunsch_lehrerin, k-abc_sed, interview, hs, legasthenieschlussbericht, diagnost_termin, wn |
| 8 | uebertritt, situation, noetig, therapi, mueh, sv_wld, einverstand, austausch, termin, pl | schlussgespraech_sr, sond, thera, schulpsycholog_abklaer, lektion_woechent, psychodiagnost, vg_abkl, th, leistungs- |
| 9 | kram, sr, ilz, wunsch, sif, motti, fortschritt, vg, gespraech, evtl | intelligenzstatus, bad, kle, kit, woechent, sv_wld, ej, mutt_abkl, ngste, audi |
| 10 | elt, info, spd, problem, schwierig, wld_agd, sed, gut, srp | sr_schlussbericht, time-out, z.h_sr, lrs-ther, textverstaendnis, rt, spezial, antragsschreib_-schulrat, thema, slp |

*Notes:* This table represents the tokens occurring with the highest probability in each of the 10 topics, as well as the tokens that are the most frequent and most exclusive in each topic. The method of topic extraction is STM, and topics are not necessarily interpretable.

Table A.1: Topics from a STM on main sample

STM and the treatment status as a content covariate. This additional reduction is used together with the STM topics to perform traditional matching. This method ensures that matched documents are similar in their topics and within-topic treatment propensity. In this study, I predict the TIRM sufficient prediction score for each treatment status (similarly to the propensity score estimation). I then use the score as additional covariate for the estimation of the nuisance parameter for the propensity score.

To give an example on how topic modeling can be used to remove confounding, I extract an STM on 10 topics with the covariates presented in Table 2.1 as topic prevalence covariates. The topic content is presented in Table A.1 (in German) and the topic distribution across treatment states is shown in Figure A.3.[86] Even though STM topics are not always directly interpretable, Table A.1 shows interesting patterns. Topic 2, for instance, seems to relate to the evaluation of learning disabilities with an IQ test and the discussion of the results with parents, whereas topic 5 reports behavioral issues and the discussion of school reports. Interestingly, topics highlight the fact that parents are active in the process (see the occurrence of the tokens "mutt" and "elt"). This is valuable information, as it allows me to account for parental influence in my estimates.

When looking at whether topics are discriminatory in terms of treatment assignment, all topics are represented in all treatments, but some seem to be more prevalent in some treatments. For instance, topics 1 and 2 are more represented in children who have been segregated in special classrooms, while topics 5 and 6 are more represented in students sent to inclusion. Even though it could be interesting to find meaning in topics, the *true* number of topics is never known, which makes topics sensitive to the ad-hoc choice of $k$ (Roberts, Stewart, and Airoldi, 2016). For this reason, I estimate STM models with different $k$.

**Word embedding with Word2Vec** The word embedding representation estimates the semantic proximity of words. The algorithm I use in this study is the Word2vec of Mikolov et al. (2013), who propose a neural network architecture to represent words in vectors as a function of their use and the words that most commonly co-occur with them. I use the *Word2Vec* embedding with word vectors of length $K$=50, 100 (I also tried with 200 and 500) based on text which is not preprocessed. For each embedding, I compute a document-level vector of length $K$ by taking the average of the numeric vectors of all the words within a document (similar to Mozer et al. (2020)). The result is an $N \times K$ matrix. Documents that are similar to each other in terms of words and their context will therefore be similar on the $K$ dimensions. Word embeddings are famous for the word associations they can produce, and I give some illustrative examples in Section A.1.2. For instance, the words that are the closest to "ADHD" and "dyslexia" (literally, "spelling disorder") are close in meaning to the two expressions. This method is therefore a good way to capture conceptual proximity between words, and accounts for the context of words.

[86] Note that this topic distribution is presented as an example. In the main analysis, it will slightly differ because of the cross-fitting strategy. However, if data are randomly cross-validated, there must be no big discrepancy between Table A.1 and the topics extracted in each fold (even though stm can

| Word | Most similar | Similarity |
|---|---|---|
| adhs | ads | 0.84 |
| adhs | adhd | 0.76 |
| adhs | neuropsycholog | 0.74 |
| adhs | pos | 0.73 |
| adhs | autismus | 0.73 |
| adhs | medizinische | 0.71 |
| adhs | erhaertet | 0.70 |
| adhs | mediz | 0.70 |
| adhs | neurologische | 0.70 |
| adhs | asperger | 0.70 |
| rechtschreibstoerung | erschwerten | 0.77 |
| rechtschreibstoerung | rezeptive | 0.76 |
| rechtschreibstoerung | beeintraechtigung | 0.75 |
| rechtschreibstoerung | auditiver | 0.74 |
| rechtschreibstoerung | sprachstoerung | 0.73 |
| rechtschreibstoerung | rechtschreibschwaeche | 0.72 |
| rechtschreibstoerung | spracherwerbsstoerung | 0.72 |
| rechtschreibstoerung | teilleistungsstaerung | 0.71 |
| rechtschreibstoerung | rechtschreibschwierigkeiten | 0.71 |
| rechtschreibstoerung | ausgepraegte | 0.71 |

Table A.2: Word2Vec word similarities

*Notes:* Illustrative examples of word associations produced by word embeddings and *Word2Vec* in German. "ADHS" is the German abbreviation for "ADHD", and "rechtschreibstoerung" is the German translation of "spelling disorder".

**Mapping mental diagnoses with *ad-hoc* dictionary approach** The text representation that is the most directly interpretable and perhaps the most convincing in this context is a dictionary approach that leverages independent mental health diagnoses. I use an independent sample on children from the City of St. Gallen, which contains an additional observed diagnosis variable given by the caseworker. I build a lexicon that, for each of the 16 formal diagnoses assigned in the City sample, identifies the tokens that are the most "key" in each diagnosis (namely, the "keywords" that are the most exclusive, or predictive, of each diagnosis). There are many ways to define "keyness", and I propose a combination of different measures that are common in computational linguistics.[87] Namely, I first take the 40 most frequent tokens per diagnosis (based on count frequency), then the 60 most frequent tokens weighted by *tf-idf*, the 40 tokens

slightly vary with changing samples).

[87] *Keyness* is the importance of a keyword within its context. To compute keyness, the frequency of a keyword in a target category for the observed frequency of a word (one particular diagnosis) is compared with its frequency in a reference category (the expected frequency, in all the other diagnoses).

with the highest chi-squared keyness value, the 40 tokens with the highest likelihood ratio G2 statistics and the 40 tokens with the highest pointwise mutual information statistics. I then take the union of all tokens provided by each measure.[88]

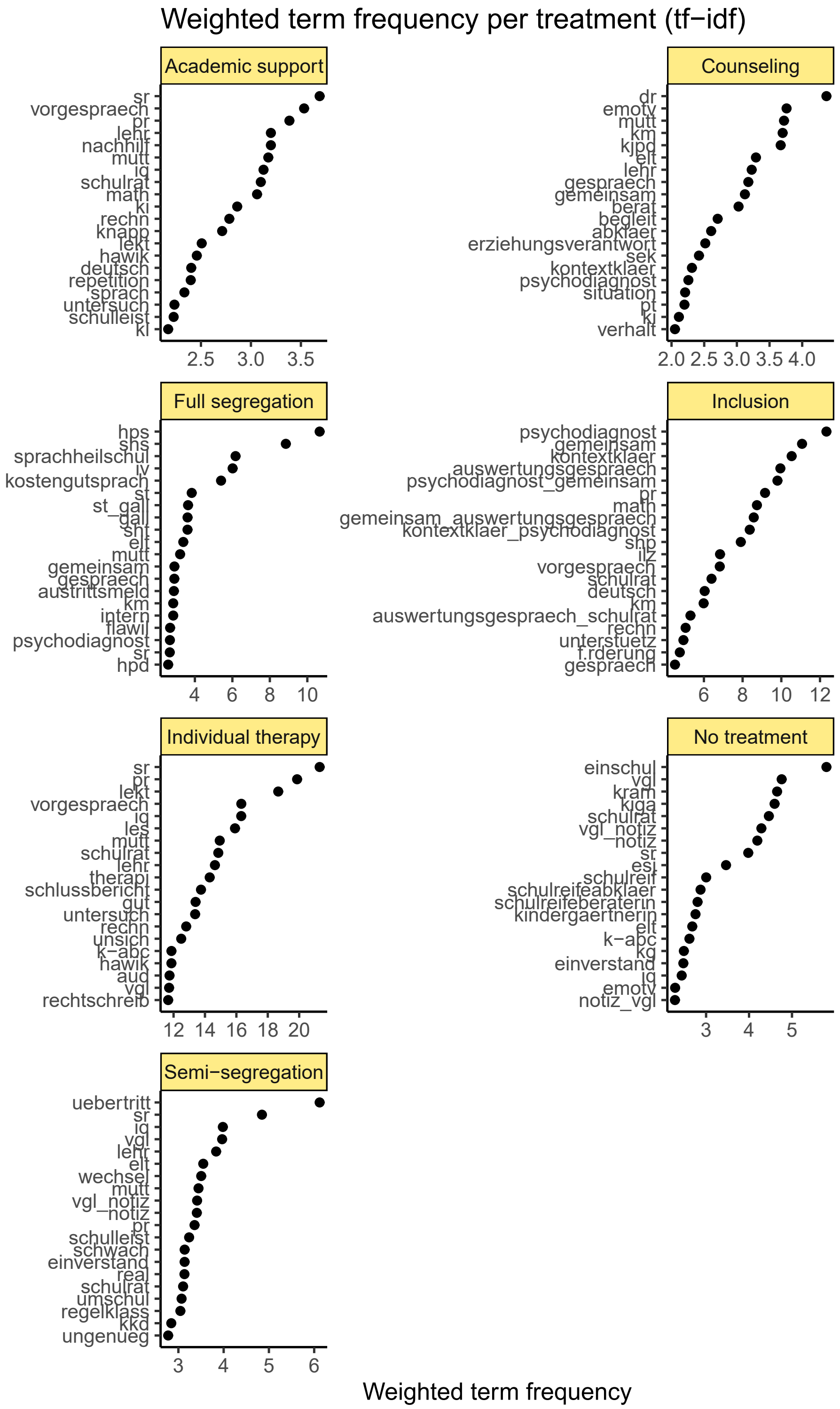
To classify documents into diagnoses, I first map each document to the diagnoses using the key tokens of each diagnosis extracted in the St. Gallen sample. I then compute the share of diagnoses per document by dividing the frequency of tokens assigned to a given diagnosis in a document by the total number of tokens per document. I weight diagnoses such that the sum of each document's diagnosis frequency sums up to 1. As a result, I obtain a vector of length 16 with the predicted proportion of diagnoses per document. In a second measure, I ensure that the most prominent diagnosis assigned to a document is discriminatory enough. I hot-encode the diagnosis if the proportion of the given diagnosis is 1.5 standard deviations above the mean of diagnosis frequency within the document.

The dictionary/keyword approach maps documents into clinically meaningful concepts by using natural language which is almost exactly similar to the language used by therapists in the Canton of St. Gallen. Therefore, it is the closest to what a human coder would do if she had to classify the caseworkers' comments. Moreover, contrary to STM and word embeddings, it is supervised. Since therapists from the City and therapists from the Canton work in the same environment, under the same rules, and use the same lexicon, texts are very similar.

Distribution of diagnoses across treatment status is presented in Figure A.4. It is interesting to notice that individual therapies, inclusion and semi-segregation share a roughly similar population of students, namely students diagnosed with problems related to school performance. Inclusion has a relatively higher share of students with learning disabilities, while semi-segregation has a higher share of students with behavioral problems. In contrast, hard segregation is particularly targeted to students with more severe disabilities, such as motor problems, and students with parental educational deficit. As I can expect, students not given any treatment display a more equal share of all diagnoses.[89]

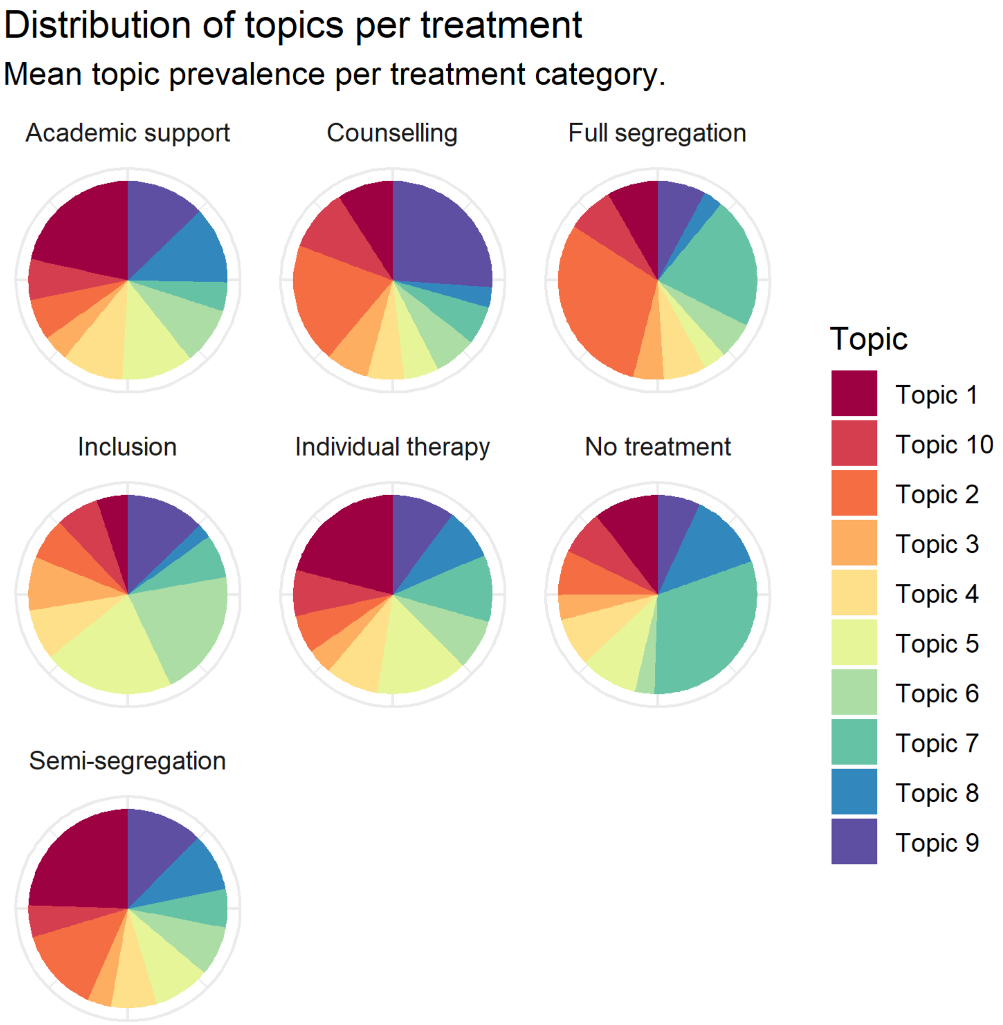
[88] Alternatively, this purely frequency analysis could be done by training a classifier on text tokens.

[89] A similar descriptive picture holds when I use 1 standard deviations above the mean to classify diagnoses.



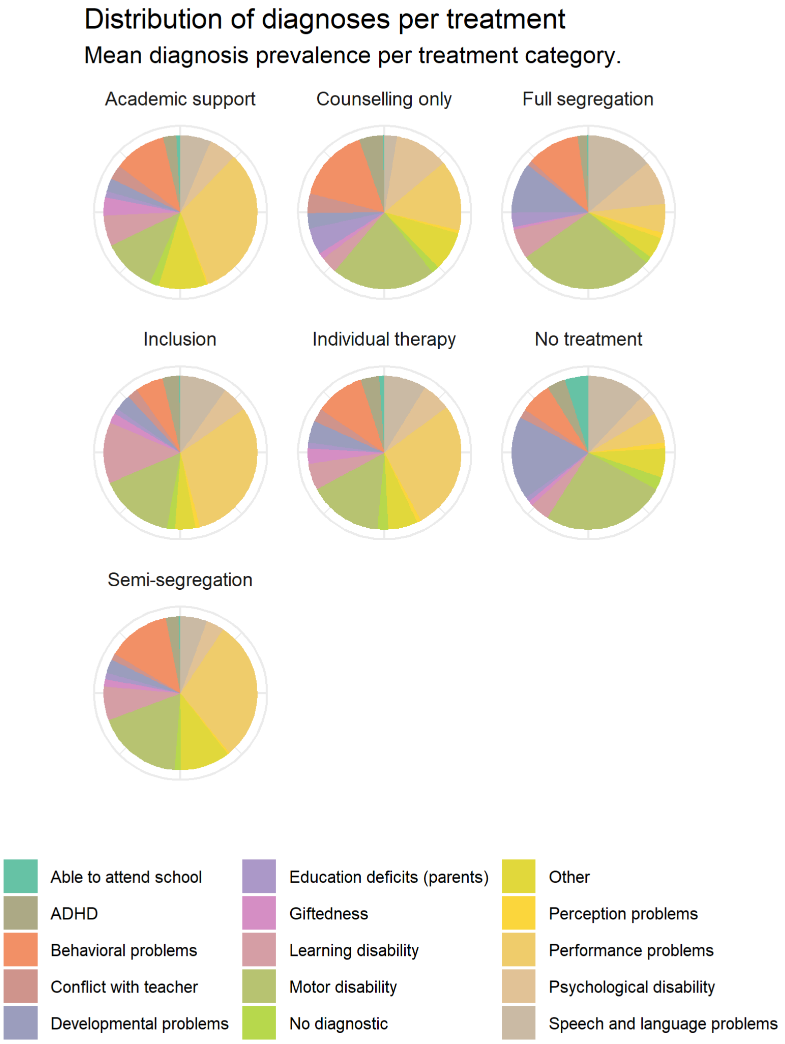
*Notes:* This figure represents the 20 most frequent terms (tokens) weighted by *tf-idf* per treatment assignment. Text is preprocessed via lemmatization. *Source: SPS.*

Figure A.2: Most frequent weighted tokens per treatment assignment



*Notes:* The prevalence of each STM topic per treatment assignment is represented. I compute the mean prevalence of each of the 10 topics per treatment category. Topic prevalence in STM is given by the main covariates. *Source: SPS*.

Figure A.3: Prevalence of STM topics per treatment assignment



*Notes:* The main 15 diagnoses extracted from an independent dataset are represented (the diagnosis "foreign language" is not displayed here). Diagnosis are extracted from the data of the City of St. Gallen *Source: SPS*.

Figure A.4: Average diagnosis per treatment assignment

## A.2 Appendix: Robustness and sensitivity checks

### A.2.1 Overlap: ATO and alternative trimming schemes

The problem of overlap is especially exacerbated in settings with many treatments. It becomes even more important with a high-dimensional, highly predictive, covariate space (see D'Amour et al., 2021). In such settings, overlap is difficult to obtain, which induces bias and extreme variability of ATEs estimates. To ensure overlap and remove extreme weights, I check the robustness of my results by estimating the Average Treatment Effect on Overlap (ATO), and by using different trimming rules.

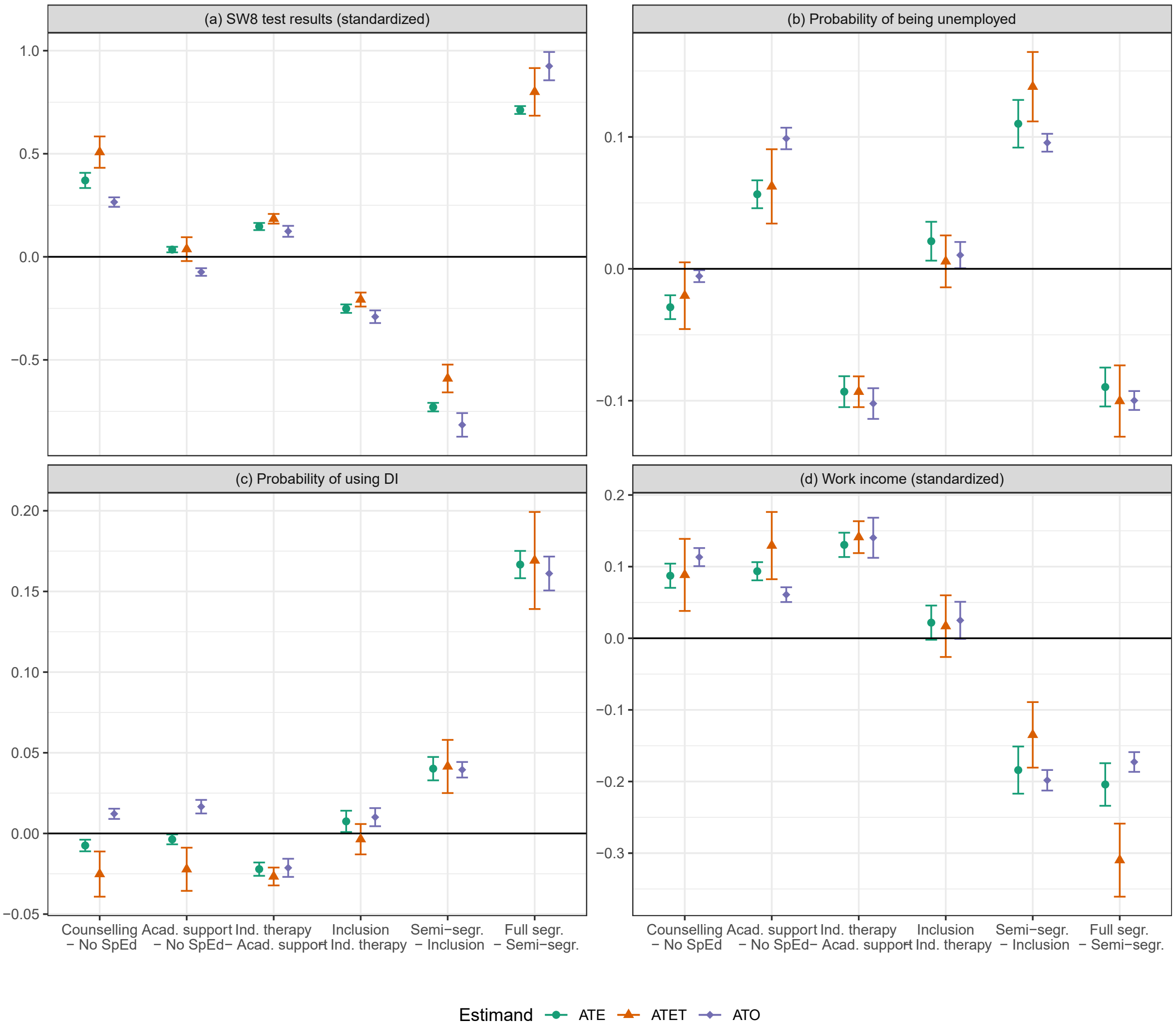
**Average treatment effect on the population of Overlap (ATO)** If students with SEN differ greatly across programs and overlap is poor, it might be relevant from a policy perspective to look at pairwise treatment effects for the population which is the most similar in terms of covariates across multiple treatments. I estimate overlap-weighted average treatment effects (Li, Morgan, and Zaslavsky, 2018; Li and Li, 2019), i.e. average treatment effects on the population of propensity score overlap $\text{ATO}_{d,d'} = E_{\text{overlap}}[Y_i^d - Y_i^{d'}]$. The ATO estimand is the following:

$$\text{ATO}_{d,d'} = E_{\text{overlap}}[\Gamma^h(d, X_i) - \Gamma^h(d', X_i)], \quad h(x) = \sum_{k=1}^{D} (\frac{1}{p_k(x)})^{-1} \tag{A.1}$$

The ATO score gives the most relative weight to the covariate regions in which none of the propensities are close to zero. It is the product of the IPW and the harmonic mean of the generalized propensity scores. Beyond focusing on an interesting population, the advantage of using the ATO is that it mitigates problems of extreme propensity scores in the ATE computation. The ATO score is not doubly-robust, and is sensitive to potential misspecifications in the propensity score.[90]

Results for the ATO are presented for each pairwise treatment effect in Figure A.1, and they are compared to the ATE and the ATET. The point estimates of the ATO are, in most cases, quite similar to the ATE point estimates, which indicates that there is good

[90] Since the weight is a function of the propensity score, it is not consistent if the propensity score is misspecified. However, as pointed out by Li and Li (2019), outcome regression may still increase the efficiency of the weighting estimator. I estimated the variance of the ATO the same way I estimated the variance of the ATE and the ATET. For more information, see Li and Li (2019) and the R-Package `psweight`.



*Notes:* This figure depicts pairwise treatment effects for Special Education programs in St. Gallen. Each pair compares interventions that are the closest in degree of severity and inclusion. Each pairwise treatment effect is the effect of being assigned to the first program in comparison to the second program on one of the four outcomes presented in the panel headers. The treatment effect on the whole population (ATE), on the population of the treated (ATET), and on the population of overlap (ATO) (see Li and Li, 2019) are presented. "Ind. therapy" is the abbreviation for individual therapies, "Acad. support" for academic support, "segr" for segregation, and "no SpEd" for receiving no program. Nuisance parameters are estimated using an ensemble learner that includes text representations presented in the "data" section. 95% confidence intervals are represented and are based on one sample $t - test$ for the ATE and the ATET. Test results and wages are standardized with mean 0 and standard deviation 1. *Source: SPS.*

Figure A.1: Pairwise treatment effects with ATO estimates

overlap in the overall population. This suggests that my ATE results do not suffer from lack of overlap.

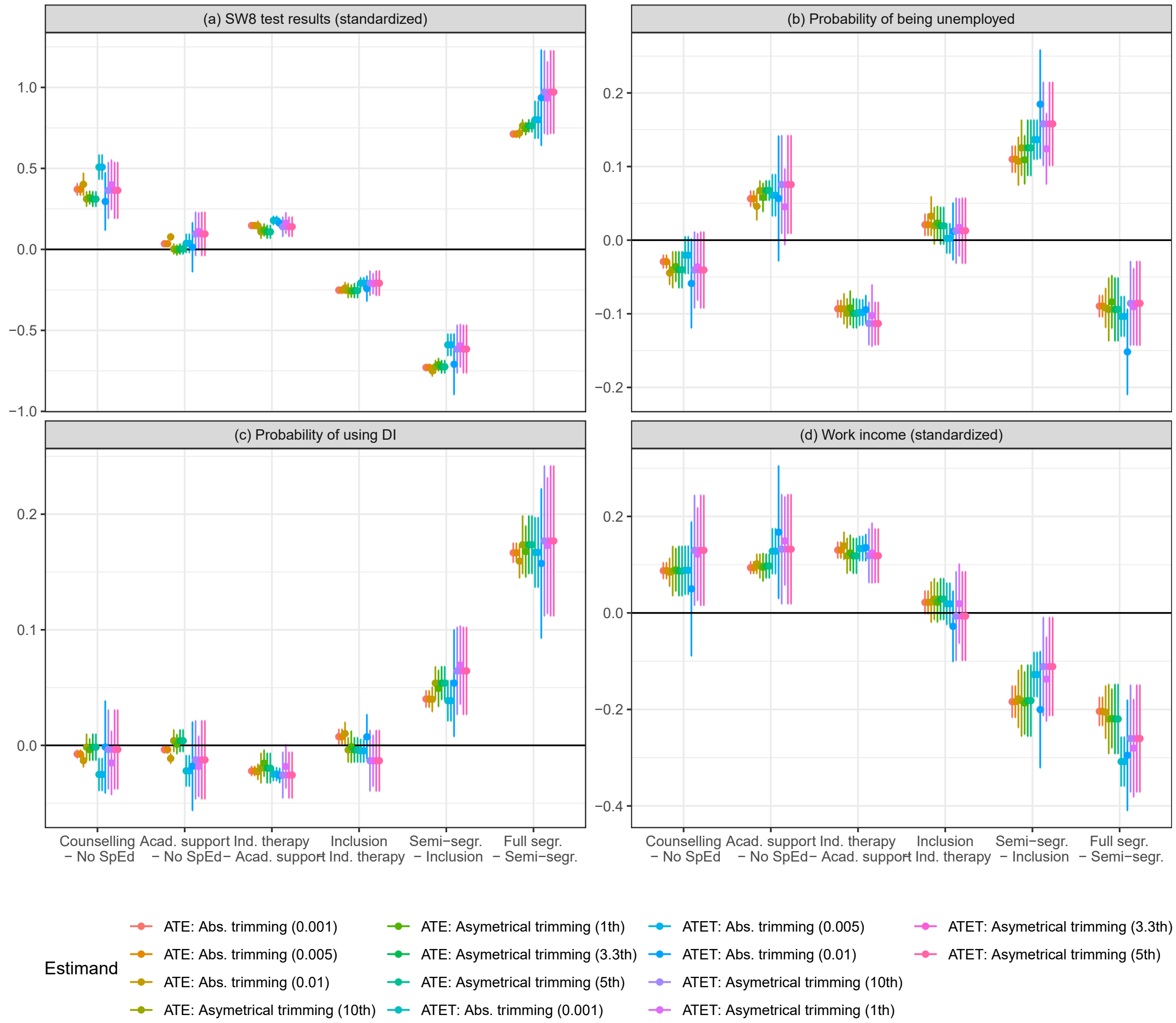
**Trimming** The idea of trimming is to remove observations with weights $h(x)$ in Equation (2.1) below a certain threshold, such that $h(x) = \underline{1}(x \in C)$ with $C$ denoting the target subpopulation defined by a threshold of the propensity score distribution $\alpha$. Each trimming rule slightly shifts the population of interest $C$. In other words, the more conservative the trimming rule, the more homogeneous the population across treatment states.

To define the trimming schemes and trimming thresholds $\alpha$, I explore different options. Following the adaptation of trimming schemes for the multi-treatment case in Yoshida et al. (2019), I trim, in a first setting, observations with all propensity scores below a certain threshold of the propensity score (see Crump et al., 2009). This referenced under "absolute trimming" with $\alpha = \{0.001, 0.005, 0.01\}$. In a second setting, the "asymmetrical trimming", I trim treated observations with their corresponding propensity score within a chosen quantile of each propensity score ("asymmetrical trimming", see Stürmer et al. (2010)). I use $\alpha = \{0.01, 0.033, 0.05, 0.1\}$ for the 1st., 3.3th, 5th and 10th quantiles. Results are presented in Figure A.2: the majority of results discussed in Section 2.4 persist across different trimming schemes. The interpretation of effect variation under different trimming rules must however be done with caution, as each trimming rule "shifts" the population of interest.

### A.2.2 Handling text as covariate

In this section, I check that my results do not depend heavily on the way I retrieve information from the text. First, I investigate the sensitivity of my estimates to the inclusion of text covariates. Second, I explore the problem of "text-induced endogeneity", which might arise if the text representation captures the psychologist's biases (towards a certain treatment or a certain writing style) rather than information on the student.

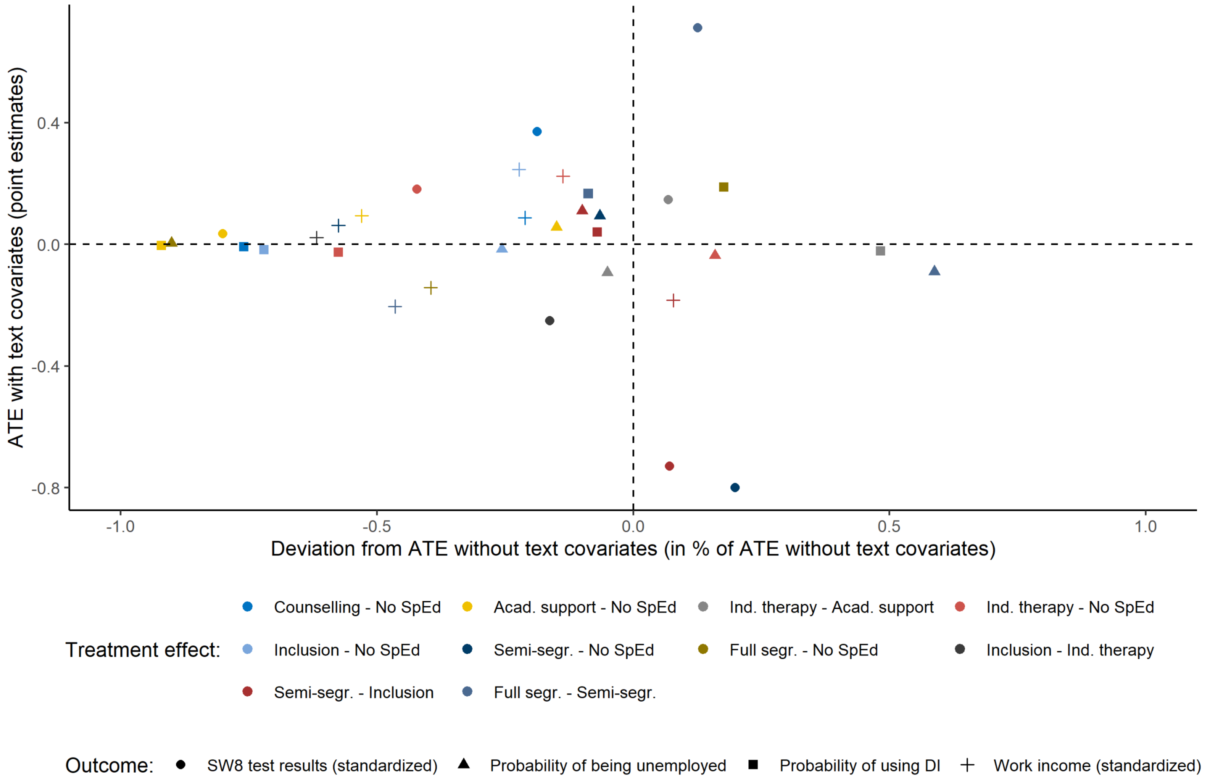
**Sensitivity to text covariates** To investigate the sensitivity of my results to the inclusion of text covariates, I compute all my results using only nontext covariates. I then show how much the estimates based on text vary with respect to estimates obtained without controlling for text. This exercise gives me an approximate indication on how much confounding I can remove by using the text.



*Notes:* This figure depicts relevant pairwise treatment effects for Special Education programs in St. Gallen at different trimming levels. Absolute trimming of Crump et al. (2009) and asymmetrical trimming of Stürmer et al. (2010) are represented at different trimming thresholds. Each pairwise treatment effect is the effect of being assigned to the first program in comparison to the second program on one of the four outcomes presented in the column headers. Both the treatment effect on the whole population (ATE) and on the population of the treated (ATET) are presented. “Ind. therapy” is the abbreviation for individual therapies, “Acad. support” for academic support, “segr.” for segregation and “no SpEd” for receiving no program. Nuisance parameters are estimated using an ensemble learner that includes text representations presented in the “data” section. 95% confidence intervals are represented and are based on one sample $t - test$ for the ATE and the ATET. *Source: SPS*.

Figure A.2: Pairwise treatment effects with different trimming rules

Figure A.3 gives the variation in effect between the specification with text covariates and the specification without text. On the $y$ axis, the ATE with text covariates is represented for each outcome and pairwise treatment effect investigated in this paper. The $x$ axis gives the difference between the ATE estimated with text covariates and the ATE estimated without text covariates in percent of the ATE estimated with-



*Notes:* This figures gives the effect variation between the specification with text covariates and the specification without text. On the $y$ axis, the ATE with text covariates is represented for each outcome and pairwise treatment effect investigated in this paper. The $x$ axis gives the difference between the ATE estimated with text covariates and the ATE estimated without text covariates in percent of the ATE estimated without text covariates.

Figure A.3: Effect comparison with specification without text covariates

out text covariates. On average, I find that estimates based on both covariates and text information are 29% smaller than estimates that do not leverage the text information. Interestingly, differences between estimates with text and estimates without text are particularly pronounced for comparisons with the "No SpEd" intervention, which suggests that there is valuable confounding information contained in the text for this particular population of students.

**Text-induced endogeneity** It is inherent to the discovery of latent features in text data that some dimensions (latent or not) of text might be exogenous to the child's characteristics but influencing treatment assignment. An example is if a given psychologist is biased towards a particular treatment, and that this psychologist always writes using a similar set of words-tokens, then the information retrieved from test will capture the psychologist's biases together with the information on the student.

To tackle this problem, I conduct two analyses. In the first analysis, I explore whether the text reflects the psychologist's writing style. I program a classifier to predict from the text the psychologist who wrote the report. If the classifier is not able to capture the psychologists' writing styles, the risk of including unwanted exogenous variation in my estimates is minimized. Running a random forest classifier to predict the caseworker, my best measure of text (frequency weighted term frequency matrix) has a prediction error of 40%, meaning that 40% of all psychologists are misclassified. However, this measure goes as low as 59% for word embeddings with 100 features and 73% for stm with 80 topics. In conclude that the psychologists do not influence the distribution of text covariates in a systematic manner.

In the second analysis, I systematically remove, for each psychologist, the most frequently and uniquely used tokens (I compute the "keyness" score across psychologists using different measures such as the chi-squared measure, likelihood ratio and point-wise mutual information). Thus, words that are too predictive of a given psychologists do not influence the estimates. I then compute my main text measures on this reduced set of features. The intuition of this procedure is somewhat similar in spirit to including psychologists' fixed effects in the regression. Results are similar to the main results presented (figures available on request).

### A.2.3 Selective Attrition

Potential selective attrition in the measured outcomes could undermine the validity of the results insofar as outcomes are not observed for all individuals. For test scores, I explicitly modeled selection into test taking and presented results above. To tackle the problem of selective attrition, I further narrow the sample to individuals for which I observe all outcomes ($N = 8993$). Estimates are presented in Figure A.4. Results are in line with main results.

### A.2.4 Exogenous placement into inclusive vs. semi-segregated Special Education programs

I turn my attention to remaining potentially unaddressed selection in the estimation of the causal effect of inclusion vs. semi-segregation. I leverage residual unexplained variation in the probability to assign students to inclusion using instrumental

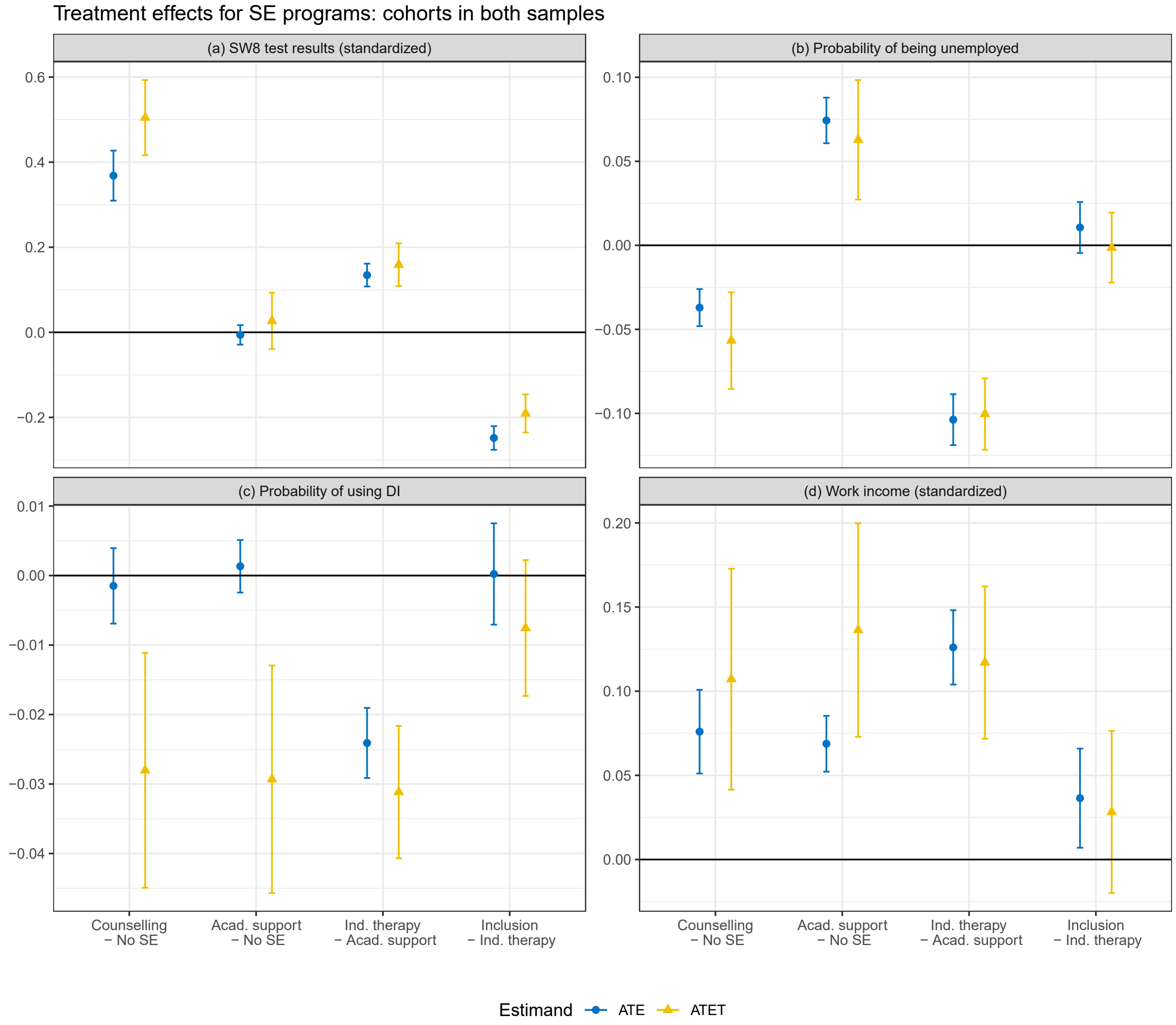


Figure A.4: Pairwise treatment effects for cohorts observed in all outcomes

This figure depicts relevant pairwise treatment effects for Special Education programs in St. Gallen. Each pairwise treatment effect is the effect of being assigned to the first program in comparison to the second program on one of the four outcomes presented in the column headers. Both the treatment effect on the whole population (ATE) and on the population of the treated (ATET) are presented. “Ind. therapy” is the abbreviation for individual therapies, “Acad. support” for academic support, “segr.” for segregation and “no SpEd” for receiving no program. Nuisance parameters are estimated using an ensemble learner that includes text representations presented in the “data” section. 95% confidence intervals are represented and are based on one sample $t - test$ for the ATE and the ATET. *Source: SPS.*

variables in the spirit of “judge-design” studies (see, e.g., Maestas, Mullen, and Strand, 2013). The main source of exogenous variation in the assignment of students to SpEd education programs and to inclusion instead of semi-segregation is the variation in the probability to assign students with SEN to inclusion *across* schools *within* years. This variation is shown in Figure 2.2. Some variation remains even when years fixed effects are introduced, and more importantly, when school-year characteristics are accounted for. Characteristics include reasons for referral, main characteristics, IQ, and whether parents decided to send the child to the SPS. School characteristics include the percentage of nonnative students, the percentage of students with SEN, the socio-economic index, the school size, and the expenditures per student.

I measure variations in the probability to assign students with SEN to inclusion *across* schools *within* years as the school-year deviations in assignment to inclusion from the mean inclusion assignment rate at the year level. This variation gives how much individual schools differ in the probability to assign inclusion from all the other schools within a year. I exploit treatment assignment deviations from mean rates of assignment to inclusive treatment as a instruments for assignment to inclusion. Preferences at the school level for inclusive treatment increase the likelihood that students with SEN will be sent to inclusive SpEd, especially for students with moderate difficulties. The key assumption that underlies my approach is that the assignment of students with SEN to a school-year is uncorrelated with unobserved characteristics (such as SEN severity) conditional on observed characteristics. As pointed out by Maestas, Mullen, and Strand (2013), this amounts to an assumption of conditional random assignment to school-year within a year.

One potential threat to this assumption is that students could select into schools. However, my rich information on school characteristics and the referral process allows me to control for this. Moreover, a careful reading of ministry reports in St. Gallen (for instance, the *Nachtrag zum Volkschulgesetz 2013*) and school documents show that schools' preferences in terms of inclusion and segregation do not, for the most case, follow a predictable pattern. Moreover, the diagnosis and treatment assignment are done in a centralized manner by independent psychologists, and they are not influenced by financial constraints at the school level. Finally, it is documented that student mobility between school is rare in the canton of St. Gallen. Families must move to another municipality if they want to change school (or enroll their students in private schools) (Balestra, Eugster, and Liebert, forthcoming, 2020; Balestra, Sallin, and Wolter, forthcoming).

In Table A.1, I present first-stage estimates and add covariates sequentially to the regression in order to indirectly test for random assignment within year on the basis of observable characteristics. The idea is that only covariates that are correlated with the deviation will affect the estimated coefficient on the deviation when included. The

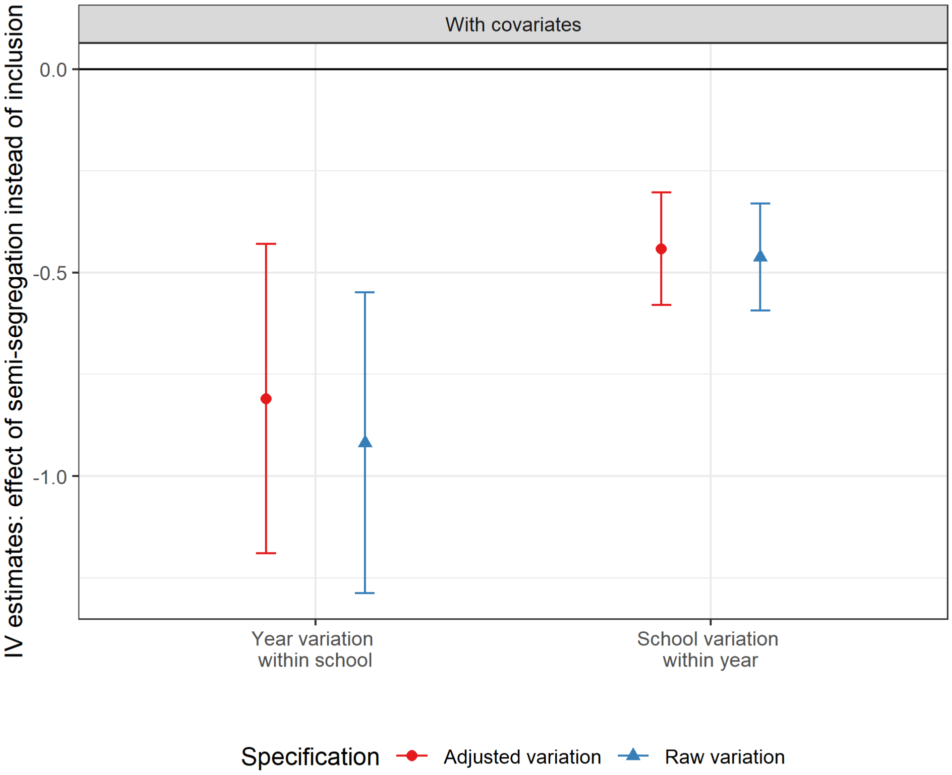
| | Probability to be assigned to inclusion | | | | |
|---|---|---|---|---|---|
| | (I) | (II) | (III) | (IV) | (V) |
| Coefficient on raw variation<br>(within year across schools) | $0.89^{***}$ | $0.87^{***}$ | $0.82^{***}$ | $0.83^{***}$ | $0.78^{***}$ |
| Coefficient on raw variation<br>(within year across schools) | $0.76^{***}$ | $0.76^{***}$ | $0.76^{***}$ | $0.76^{***}$ | $0.75^{***}$ |
| Coefficient on raw variation<br>(within school across years) | $0.44^{***}$ | $0.48^{***}$ | $0.45^{***}$ | $0.44^{***}$ | $0.45^{***}$ |
| Coefficient on adjusted variation<br>(within school across years) | $0.37^{***}$ | $0.38^{***}$ | $0.39^{***}$ | $0.39^{***}$ | $0.40^{***}$ |
| **Control variables included** | | | | | |
| Students' characteristics | | X | X | X | X |
| IQ score | | | X | X | X |
| Reasons for referral | | | | X | X |
| School characteristics per year | | | | | X |

*Notes:* First stage regression of the schools preferences for inclusion on the assignment to inclusion. All regressions include schools fixed effects for the regressions within schools, and year fixed effects for the regressions within years. Number of observations: 2932 for raw variation and 2497 for adjusted variation. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.001$. *Source: SPS*

Table A.1: First-Stage Regressions: Effect of preference for inclusion on assignment to inclusion

coefficient on the variation is not significantly affected by the addition of variables, not even by the addition of school-year characteristics such as socio-economic status of the school or by the IQ score. This is the case both when the raw and the adjusted variation is considered. Thus, these results show that the measured variations are not correlated with student or school characteristics, and thus the preference for inclusion depends on schools' preferences. Note that, since, in a multiple treatment setting, a sound IV approach requires one instrument per treatment (see Kirkeboen, Leuven, and Mogstad, 2016), I focus only on SN students having received either inclusion (excluding individual therapies) or semi-segregation and restrict my dataset accordingly. This also ensures that I measure the effect for a homogeneous population of students with SN (students who have issues such that they would either assigned to inclusion or semi-segregation). However, the first stage estimates hold also when the whole sample is considered.

I estimate (Conditional) Local Average Treatment Effects of being assigned to inclusion rather than segregation on test scores using unexplained variation in the decision to implement inclusion by school within years or by year within school. Similar to



*Notes:* This figure shows LATE estimates of the effect of inclusion for academic performance and unemployment probability for students either in inclusive or segregated settings. The estimates are found with 2SLS, and covariates include student and school characteristics. 95% confidence intervals are represented. Both the effect for the adjusted and the raw deviations in the first stage are represented.

Figure A.5: LATE for inclusion vs. semi-segregation

Maestas, Mullen, and Strand (2013), I estimate 2SLS with and without covariates, both using the raw and the adjusted variation. I do not use text information in this case. Results are presented in Figure A.5. LATE estimates show that semi-segregation has a negative impact on students' academic performance. The impact of semi-segregation is around -0.45 standard deviations of test score in comparison to inclusion, and it is of 10 percentage points on the probability to be unemployed. The fact that the test score estimates are substantially lower than the ATE estimates is mostly due to the fact that the populations of interest are not exactly the same: ATEs measure the effect for the whole population of students with SEN, whereas the LATE measures the effect on compliers, i.e. students who would have been assigned to semi-segregation if the school did not have a "preference" for inclusion.

In conclusion, this IV exercise allowed me to exploit some existing exogenous variation in treatment assignment across schools within years to estimate treatment effects. It is, however, not the panacea. First, potential difficulties with the monotonicity might persist when more than two treatments exist. Even though I restricted my analysis to the subsample of students observed in inclusion and semi-segregation, the instrument could still move some students to other available treatments. Second, even though I observe many characteristics about schools, the (almost unsolvable) issue of not being able to observe the actual teaching behaviors and teaching styles of both the main teachers and the SpEd teachers at the school level still leaves the black box open.

## A.3 Appendix: Supplementary Material

| Therapy | German | Treatment group |
|---|---|---|
| Inclusive special education (ISF) | Integrierte Schülerförderung (ISF),<br>Schulische Heilpädagogik | Inclusion |
| Special (small) classes | Kleinklasse | Semi-segregation |
| Speech therapy | Logotherapie | Individual therapy |
| Psychomotor therapy | Psychomotoriktherapie | Individual therapy |
| Dyslexia therapy | Legasthenie (Lese- und Rechtschreibstörung) | Individual therapy |
| Dyscalculia therapy | Dyskalkulietherapie | Individual therapy |
| Rhythm therapy (Dalcroze eurhythmics) | Rhythmik | Individual therapy |
| Tutoring, language tutoring | Hilfe | Academic support |
| Counseling | Psychologische Hilfe, Beratung | Counseling |

Table A.1: List of available therapies from the Cantonal offer

*Source*: "Sonderpädagogikkonzept für die Regelschule" from the Ministry of Education, 18.3.2015, retrieved on the official website of the Ministry of Education of St. Gallen, https://www.sg.ch/bildung-sport/volksschule/rahmenbedingungen/rechtliche-grundlagen/konzepte.html.

| *Data restrictions* | Number of observations |
|---|---|
| Full sample of students in contact with SPS | 28584 |
| - Trim cohorts (1982 to 2003) and missing birthdates | -972 |
| - Native language non-imputable | -1288 |
| - Treatment not in cantonal offer | -8612 |
| - Treatment not identified | -478 |
| **TOTAL** | **17822** |
| IQ not computed | 4801 |

Table A.2: Attrition analysis

| | Counselling | Academic support | Individual therapy | Inclusive special ed.<br>(ISF) | Semi-segregation | Full segregation | No therapy<br>(but sent to SPS) |
|---|---|---|---|---|---|---|---|
| $N$ ($N = 17,822$) | 1,450 | 1,381 | 7,997 | 2,705 | 1,690 | 1,603 | 996 |
| **A: Individual characteristics** | | | | | | | |
| Female | 0.34 | 0.50 | 0.40 | 0.46 | 0.43 | 0.30 | 0.38 |
| Foreign language | 0.08 | 0.15 | 0.09 | 0.11 | 0.28 | 0.14 | 0.20 |
| IQ | 101.47 (13.08) | 94.24 (10.54) | 98.26 (10.58) | 92.94 (9.41) | 86.17 (8.96) | 87.01 (14.99) | 93.63 (11.25) |
| IQ measured | 0.62 | 0.73 | 0.74 | 0.80 | 0.78 | 0.67 | 0.61 |
| Birth year | 1994.43 (4.53) | 1993.86 (4.25) | 1995.14 (4.33) | 1996.68 (3.73) | 1993.61 (3.64) | 1995.80 (4.31) | 1999.18 (3.41) |
| Age at first interview | 9.12 (285) | 9.47 (2.29) | 8.69 (1.96) | 8.75 (2.03) | 9.11 (2.50) | 7.19 (2.63) | 6.24 (0.60) |
| Had bridge year | 0.08 | 0.08 | 0.08 | 0.07 | 0.13 | 0.08 | 1.00 |
| Reasons: other | 0.04 | 0.03 | 0.03 | 0.02 | 0.06 | 0.10 | 0.10 |
| Reasons: social and emotional problems | 0.48 | 0.19 | 0.15 | 0.18 | 0.22 | 0.28 | 0.24 |
| Reasons: performance and learning problems | 0.68 | 0.91 | 0.92 | 0.94 | 0.88 | 0.78 | 0.87 |
| Reasons: problems with teachers or school | 0.06 | 0.02 | 0.01 | 0.05 | 0.02 | 0.04 | 0.03 |
| Reasons: not specified | 0.03 | 0.01 | 0.01 | 0.00 | 0.01 | 0.02 | 0.00 |
| Referred by: Caseworker | 0.00 | 0.01 | 0.04 | 0.02 | 0.01 | 0.05 | 0.01 |
| Referred by: Other | 0.04 | 0.02 | 0.02 | 0.01 | 0.03 | 0.07 | 0.02 |
| Referred by: Parents | 0.15 | 0.06 | 0.05 | 0.04 | 0.02 | 0.06 | 0.03 |
| Referred by: Parents and teacher | 0.62 | 0.66 | 0.71 | 0.64 | 0.61 | 0.53 | 0.58 |
| Referred by: Teacher | 0.19 | 0.25 | 0.18 | 0.28 | 0.33 | 0.29 | 0.36 |
| Total number of SPS visits | 10.69 (8.08) | 8.77 (5.94) | 9.10 (6.18) | 10.27 (7.72) | 13.58 (9.25) | 19.62 (14.79) | 6.14 (3.93) |
| Regional office: Ro | 0.12 | 0.22 | 0.15 | 0.04 | 0.15 | 0.14 | 0.15 |
| Regional office: Go | 0.09 | 0.07 | 0.13 | 0.02 | 0.12 | 0.12 | 0.04 |
| Regional office: Wi | 0.15 | 0.12 | 0.18 | 0.12 | 0.25 | 0.13 | 0.12 |
| Regional office: Wa | 0.22 | 0.10 | 0.12 | 0.17 | 0.07 | 0.19 | 0.11 |
| Regional office: Ra | 0.12 | 0.09 | 0.07 | 0.38 | 0.06 | 0.20 | 0.13 |
| Regional office: Sa | 0.23 | 0.16 | 0.18 | 0.20 | 0.18 | 0.15 | 0.20 |
| Regional office: Re | 0.07 | 0.24 | 0.17 | 0.06 | 0.18 | 0.08 | 0.25 |
| School: percent nonnatives | 0.22 (0.10) | 0.21 (0.10) | 0.22 (0.10) | 0.20 (0.09) | 0.26 (0.10) | 0.22 (0.11) | 0.24 (0.10) |
| School: percent SEN students | 0.18 (0.10) | 0.18 (0.09) | 0.18 (0.10) | 0.18 (0.11) | 0.18 (0.10) | 0.18 (0.11) | 0.19 (0.11) |
| School: social index | 0.98 (0.07) | 0.98 (0.07) | 0.98 (0.07) | 0.95 (0.06) | 1.01 (0.07) | 0.98 (0.07) | 0.99 (0.07) |
| School: total number of students | 176.37 (158.92) | 151.55 (123.93) | 161.58 (133.71) | 174.97 (194.00) | 211.08 (161.88) | 173.50 (150.32) | 165.58 (139.30) |
| School: expenditures per student (2017) | -0.01 (0.95) | 0.06 (1.05) | -0.05 (0.97) | 0.28 (1.03) | -0.23 (0.88) | 0.03 (1.10) | 0.01 (1.05) |
| School: urban | 0.48 (0.50) | 0.43 (0.50) | 0.45 (0.50) | 0.37 (0.48) | 0.60 (0.49) | 0.47 (0.50) | 0.44 (0.50) |
| **B: Sample attrition** | | | | | | | |
| In a SW8 cohort | 0.72 | 0.67 | 0.75 | 0.89 | 0.69 | 0.82 | 0.99 |
| In AHV data | 0.73 | 0.76 | 0.70 | 0.58 | 0.81 | 0.64 | 0.33 |
| In both SW8 and AHV data | 0.47 | 0.46 | 0.46 | 0.47 | 0.53 | 0.47 | 0.32 |
| **C: Outcomes** | | | | | | | |
| SW8 Test taken (in SW8 cohort) | 0.74 | 0.81 | 0.82 | 0.86 | 0.80 | 0.41 | 0.63 |
| SW8 Math and German (std) | 0.57 (1.05) | -0.09 (0.86) | 0.24 (0.88) | -0.24 (0.85) | -1.12 (0.92) | -0.20 (1.12) | 0.17 (0.95) |
| Used disability insurance | 0.06 | 0.05 | 0.03 | 0.04 | 0.11 | 0.38 | 0.07 |
| Used unemployment insurance | 0.24 | 0.32 | 0.19 | 0.19 | 0.39 | 0.25 | 0.19 |
| Last registered yearly wage (std., SSA cohort) | -0.09 (1.03) | -0.06 (0.96) | 0.15 (0.97) | 0.12 (0.98) | -0.16 (1.00) | -0.57 (0.95) | -0.15 (0.95) |

Table A.3: Summary statistics per treatment group

Summary statistics for the population of students referred to the SPS in the Canton of St. Gallen. The names of Regional offices are abbreviated for confidentiality purposes. The sample is composed of SN students from the Canton of St. Gallen having visited the SPS between 1998 and 2010 and being born between 1982 and 2003. Mean per treatment groups are reported, and standard deviations are reported in parentheses for continuous variables. *Source: SPS*

| | Average | Counseling vs Acad. support | Counseling vs Ind. ther. | Counseling vs ISF | Counseling vs Semi-segr. | Counseling vs Full segr. | Counseling vs No treatment | Acad. support vs Ind. ther. | Acad. support vs ISF | Acad. support vs Semi-segr. | Acad. support vs Full segr. | Acad. support vs No treatment | Ind. ther. vs ISF | Ind. ther. vs Semi-segr. | Ind. ther. vs Full segr. | Ind. ther. vs No treatment | ISF vs Semi-segr. | ISF vs Full segr. | ISF vs No treatment | Semi-segr. vs Full segr. | Semi-segr. vs No treatment | Full segr. vs No treatment |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Female | 0.17 | 0.33 | 0.13 | 0.26 | 0.19 | 0.08 | 0.09 | 0.20 | 0.08 | 0.14 | 0.41 | 0.24 | 0.12 | 0.05 | 0.21 | 0.04 | 0.07 | 0.34 | 0.16 | 0.27 | 0.09 | 0.17 |
| Foreign language | 0.23 | 0.23 | 0.04 | 0.11 | 0.53 | 0.19 | 0.37 | 0.19 | 0.12 | 0.30 | 0.04 | 0.14 | 0.07 | 0.49 | 0.15 | 0.33 | 0.42 | 0.07 | 0.25 | 0.35 | 0.17 | 0.18 |
| IQ | 0.58 | 0.61 | 0.27 | 0.75 | 1.37 | 1.03 | 0.64 | 0.38 | 0.13 | 0.82 | 0.56 | 0.06 | 0.53 | 1.23 | 0.87 | 0.42 | 0.74 | 0.47 | 0.07 | 0.07 | 0.73 | 0.50 |
| IQ measured | 0.20 | 0.24 | 0.27 | 0.40 | 0.36 | 0.10 | 0.01 | 0.03 | 0.16 | 0.12 | 0.14 | 0.25 | 0.12 | 0.09 | 0.17 | 0.29 | 0.03 | 0.30 | 0.41 | 0.26 | 0.38 | 0.11 |
| Birth year | 0.58 | 0.13 | 0.16 | 0.54 | 0.20 | 0.31 | 1.19 | 0.30 | 0.71 | 0.06 | 0.45 | 1.38 | 0.38 | 0.38 | 0.15 | 1.04 | 0.83 | 0.22 | 0.70 | 0.55 | 1.58 | 0.87 |
| Age at first interview | 0.67 | 0.13 | 0.18 | 0.15 | 0.01 | 0.71 | 1.40 | 0.36 | 0.33 | 0.15 | 0.92 | 1.92 | 0.03 | 0.19 | 0.65 | 1.69 | 0.16 | 0.66 | 1.67 | 0.75 | 1.57 | 0.49 |
| Had bridge year | 1.36 | 0.02 | 0.01 | 0.05 | 0.16 | 0.00 | 4.64 | 0.01 | 0.04 | 0.18 | 0.02 | 4.79 | 0.04 | 0.17 | 0.01 | 4.71 | 0.21 | 0.06 | 5.11 | 0.16 | 3.59 | 4.61 |
| Reasons: other | 0.16 | 0.07 | 0.08 | 0.12 | 0.06 | 0.22 | 0.20 | 0.01 | 0.05 | 0.13 | 0.29 | 0.27 | 0.04 | 0.13 | 0.29 | 0.27 | 0.17 | 0.33 | 0.31 | 0.17 | 0.15 | 0.02 |
| Reasons: social and emotional problems | 0.27 | 0.64 | 0.75 | 0.66 | 0.57 | 0.43 | 0.51 | 0.11 | 0.02 | 0.06 | 0.20 | 0.12 | 0.09 | 0.17 | 0.31 | 0.23 | 0.09 | 0.22 | 0.15 | 0.13 | 0.06 | 0.07 |
| Reasons: performance and learning problems | 0.30 | 0.59 | 0.64 | 0.71 | 0.50 | 0.22 | 0.46 | 0.04 | 0.13 | 0.09 | 0.37 | 0.14 | 0.08 | 0.14 | 0.41 | 0.18 | 0.22 | 0.49 | 0.26 | 0.27 | 0.04 | 0.23 |
| Reasons: problems with teachers or school | 0.11 | 0.20 | 0.22 | 0.01 | 0.17 | 0.07 | 0.15 | 0.03 | 0.19 | 0.03 | 0.13 | 0.05 | 0.22 | 0.06 | 0.16 | 0.08 | 0.17 | 0.06 | 0.15 | 0.11 | 0.02 | 0.09 |
| Reasons: not specified | 0.11 | 0.10 | 0.11 | 0.22 | 0.11 | 0.07 | 0.23 | 0.01 | 0.14 | 0.01 | 0.03 | 0.16 | 0.14 | 0.00 | 0.04 | 0.15 | 0.14 | 0.17 | 0.04 | 0.04 | 0.15 | 0.18 |
| Referred by: Caseworker | 0.14 | 0.06 | 0.25 | 0.16 | 0.06 | 0.29 | 0.03 | 0.20 | 0.11 | 0.01 | 0.25 | 0.03 | 0.10 | 0.20 | 0.06 | 0.23 | 0.11 | 0.15 | 0.14 | 0.26 | 0.03 | 0.28 |
| Referred by: Other | 0.09 | 0.09 | 0.07 | 0.14 | 0.11 | 0.09 | 0.10 | 0.02 | 0.05 | 0.03 | 0.17 | 0.01 | 0.08 | 0.05 | 0.15 | 0.03 | 0.03 | 0.22 | 0.04 | 0.20 | 0.01 | 0.18 |
| Referred by: Parents | 0.17 | 0.32 | 0.37 | 0.40 | 0.46 | 0.29 | 0.42 | 0.05 | 0.09 | 0.15 | 0.03 | 0.10 | 0.04 | 0.11 | 0.08 | 0.06 | 0.07 | 0.12 | 0.02 | 0.19 | 0.05 | 0.14 |
| Referred by: Parents and teacher | 0.13 | 0.08 | 0.18 | 0.04 | 0.01 | 0.16 | 0.08 | 0.10 | 0.04 | 0.08 | 0.24 | 0.16 | 0.14 | 0.19 | 0.35 | 0.26 | 0.05 | 0.20 | 0.12 | 0.16 | 0.07 | 0.08 |
| Referred by: Teacher | 0.18 | 0.16 | 0.01 | 0.22 | 0.33 | 0.25 | 0.38 | 0.18 | 0.06 | 0.16 | 0.09 | 0.22 | 0.24 | 0.34 | 0.26 | 0.40 | 0.10 | 0.02 | 0.16 | 0.08 | 0.06 | 0.13 |
| Referred by: NA | 0.10 | 0.02 | 0.11 | 0.11 | 0.04 | 0.12 | 0.04 | 0.09 | 0.09 | 0.06 | 0.14 | 0.02 | 0.00 | 0.14 | 0.21 | 0.07 | 0.14 | 0.22 | 0.07 | 0.09 | 0.08 | 0.16 |
| Total number of SPS visits | 0.55 | 0.27 | 0.22 | 0.05 | 0.33 | 0.75 | 0.72 | 0.06 | 0.22 | 0.62 | 0.96 | 0.52 | 0.17 | 0.57 | 0.93 | 0.57 | 0.39 | 0.79 | 0.68 | 0.49 | 1.05 | 1.25 |
| Regional office: Ro | 0.17 | 0.26 | 0.07 | 0.30 | 0.07 | 0.06 | 0.07 | 0.19 | 0.55 | 0.19 | 0.20 | 0.19 | 0.37 | 0.00 | 0.01 | 0.00 | 0.37 | 0.36 | 0.37 | 0.01 | 0.01 | 0.02 |
| Regional office: Go | 0.19 | 0.05 | 0.13 | 0.28 | 0.11 | 0.09 | 0.19 | 0.19 | 0.23 | 0.16 | 0.14 | 0.13 | 0.40 | 0.02 | 0.04 | 0.31 | 0.38 | 0.36 | 0.10 | 0.02 | 0.29 | 0.27 |
| Regional office: Wi | 0.14 | 0.09 | 0.09 | 0.08 | 0.25 | 0.06 | 0.07 | 0.19 | 0.01 | 0.34 | 0.04 | 0.02 | 0.17 | 0.16 | 0.15 | 0.16 | 0.33 | 0.02 | 0.01 | 0.30 | 0.32 | 0.01 |
| Regional office: Wa | 0.19 | 0.34 | 0.26 | 0.12 | 0.43 | 0.09 | 0.30 | 0.09 | 0.22 | 0.09 | 0.26 | 0.04 | 0.14 | 0.18 | 0.17 | 0.05 | 0.31 | 0.03 | 0.18 | 0.35 | 0.13 | 0.22 |
| Regional office: Ra | 0.32 | 0.07 | 0.14 | 0.65 | 0.19 | 0.23 | 0.05 | 0.07 | 0.72 | 0.12 | 0.30 | 0.13 | 0.79 | 0.05 | 0.37 | 0.20 | 0.84 | 0.42 | 0.60 | 0.42 | 0.24 | 0.18 |
| Regional office: Sa | 0.08 | 0.17 | 0.13 | 0.09 | 0.14 | 0.21 | 0.08 | 0.04 | 0.09 | 0.03 | 0.04 | 0.09 | 0.05 | 0.00 | 0.08 | 0.05 | 0.05 | 0.13 | 0.00 | 0.07 | 0.06 | 0.13 |
| Regional office: Re | 0.27 | 0.46 | 0.29 | 0.04 | 0.32 | 0.03 | 0.49 | 0.18 | 0.50 | 0.15 | 0.44 | 0.02 | 0.33 | 0.03 | 0.26 | 0.20 | 0.36 | 0.07 | 0.53 | 0.29 | 0.17 | 0.46 |
| In a SW8 cohort | 0.39 | 0.11 | 0.07 | 0.43 | 0.07 | 0.24 | 0.83 | 0.18 | 0.54 | 0.04 | 0.35 | 0.94 | 0.35 | 0.14 | 0.17 | 0.76 | 0.49 | 0.18 | 0.44 | 0.31 | 0.89 | 0.60 |
| In AHV data | 0.40 | 0.06 | 0.08 | 0.33 | 0.19 | 0.21 | 0.89 | 0.14 | 0.39 | 0.13 | 0.27 | 0.97 | 0.25 | 0.27 | 0.13 | 0.80 | 0.53 | 0.12 | 0.53 | 0.40 | 1.13 | 0.65 |
| In both SW8 and AHV data | 0.13 | 0.03 | 0.02 | 0.00 | 0.12 | 0.01 | 0.32 | 0.01 | 0.03 | 0.14 | 0.02 | 0.30 | 0.02 | 0.14 | 0.01 | 0.30 | 0.12 | 0.01 | 0.32 | 0.13 | 0.45 | 0.32 |
| SW8 Test taken (in SW8 cohort) | 0.39 | 0.17 | 0.19 | 0.30 | 0.14 | 0.70 | 0.22 | 0.02 | 0.13 | 0.03 | 0.89 | 0.39 | 0.11 | 0.05 | 0.92 | 0.42 | 0.16 | 1.05 | 0.53 | 0.86 | 0.36 | 0.46 |
| SW8 Math and German (std) | 0.64 | 0.68 | 0.34 | 0.85 | 1.71 | 0.71 | 0.40 | 0.38 | 0.18 | 1.15 | 0.11 | 0.28 | 0.56 | 1.52 | 0.44 | 0.08 | 0.99 | 0.04 | 0.46 | 0.90 | 1.38 | 0.35 |
| Used disability insurance | 0.34 | 0.02 | 0.12 | 0.10 | 0.19 | 0.84 | 0.05 | 0.10 | 0.08 | 0.21 | 0.86 | 0.07 | 0.02 | 0.31 | 0.94 | 0.17 | 0.29 | 0.93 | 0.15 | 0.65 | 0.15 | 0.80 |
| Used unemployment insurance | 0.20 | 0.18 | 0.12 | 0.13 | 0.33 | 0.01 | 0.12 | 0.30 | 0.31 | 0.15 | 0.17 | 0.30 | 0.01 | 0.45 | 0.13 | 0.00 | 0.46 | 0.14 | 0.01 | 0.32 | 0.45 | 0.13 |
| Income (std.) | 0.16 | 0.01 | 0.05 | 0.07 | 0.10 | 0.32 | 0.09 | 0.07 | 0.08 | 0.10 | 0.34 | 0.09 | 0.02 | 0.17 | 0.41 | 0.16 | 0.18 | 0.40 | 0.17 | 0.26 | 0.00 | 0.24 |

Table A.4: Summary statistics: SMD comparison

Standardized mean differences (SMD) across SE programs. For two SE programs $w$ and $w'$, SMDs are computed as $\frac{\bar{x}_w - \bar{x}_{w'}}{\sqrt{\frac{s_w^2 + s_{w'}^2}{2}}}$, where $\bar{x}_w$ is the mean of the covariate in treatment group $w$ and $s_w^2$ is the sample variance of covariate in treatment group $w$. A SMD above 0.2 is considered as an important difference across groups. "Acad. support" is academic support, ISF is inclusion, Semi-segr. is semi-segregation. *Source: SPS*

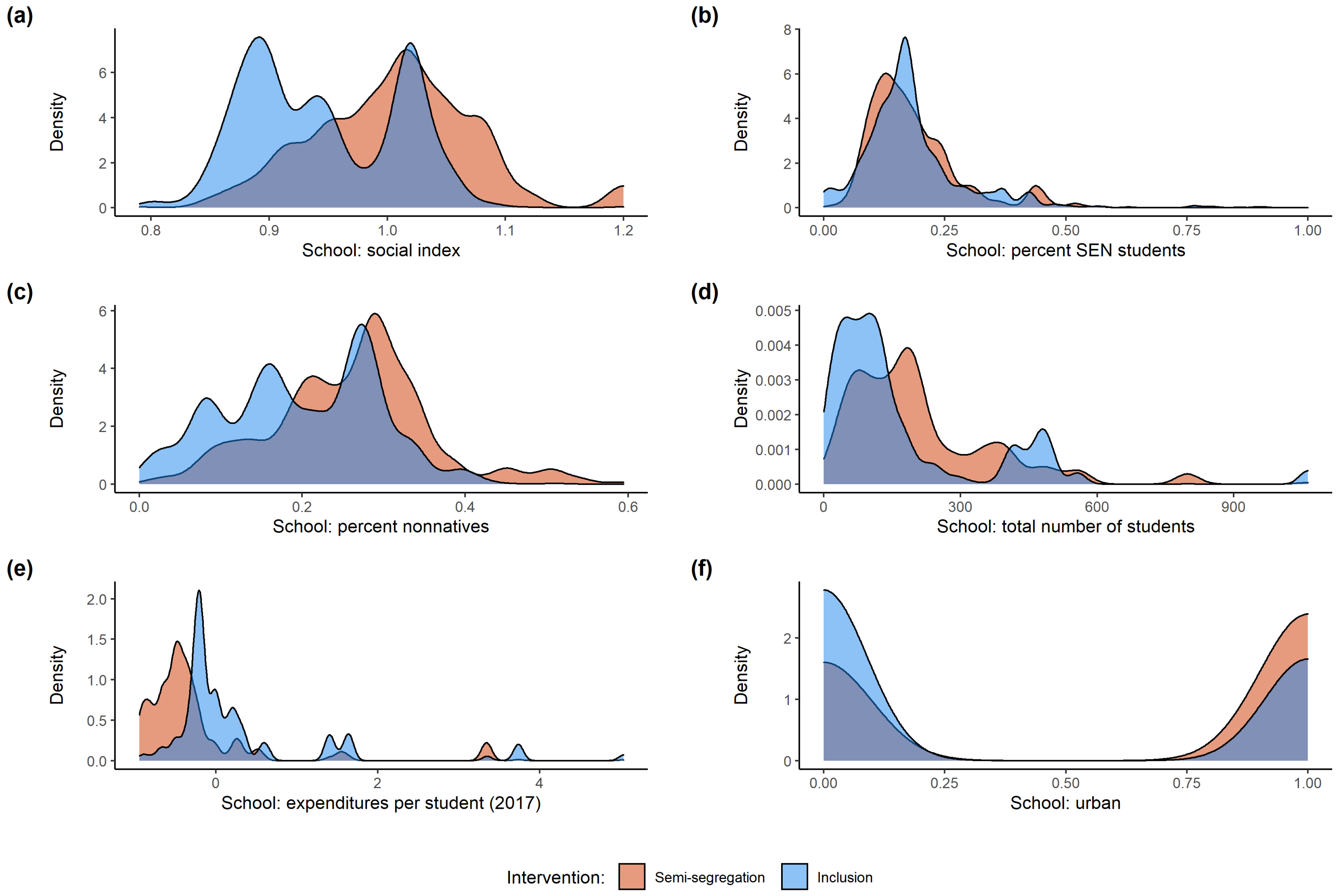
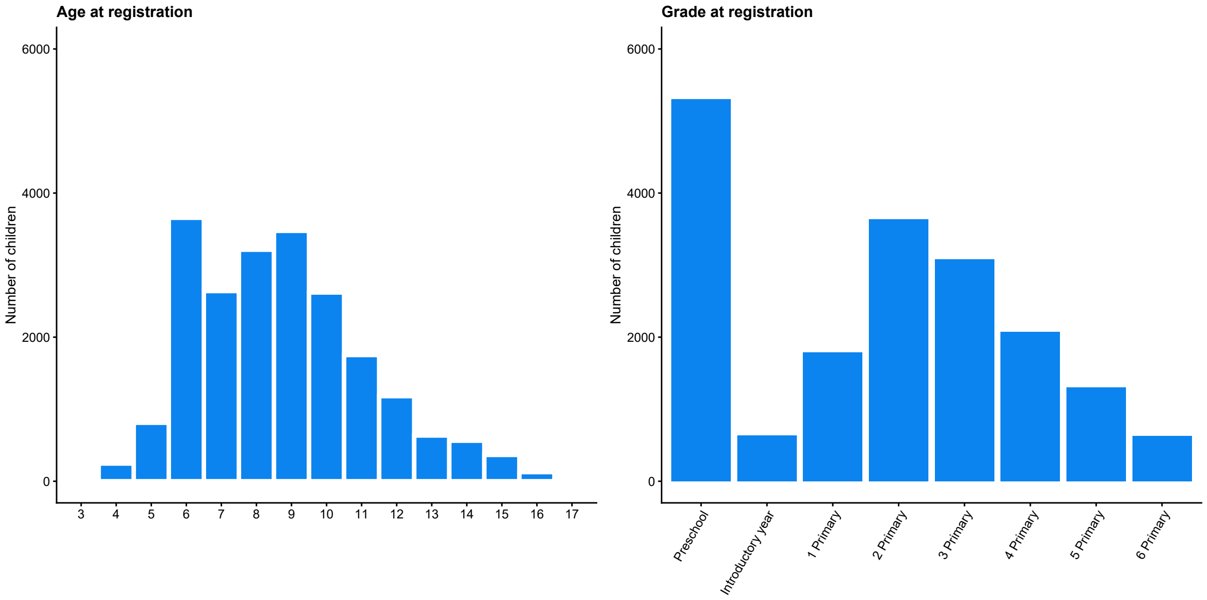


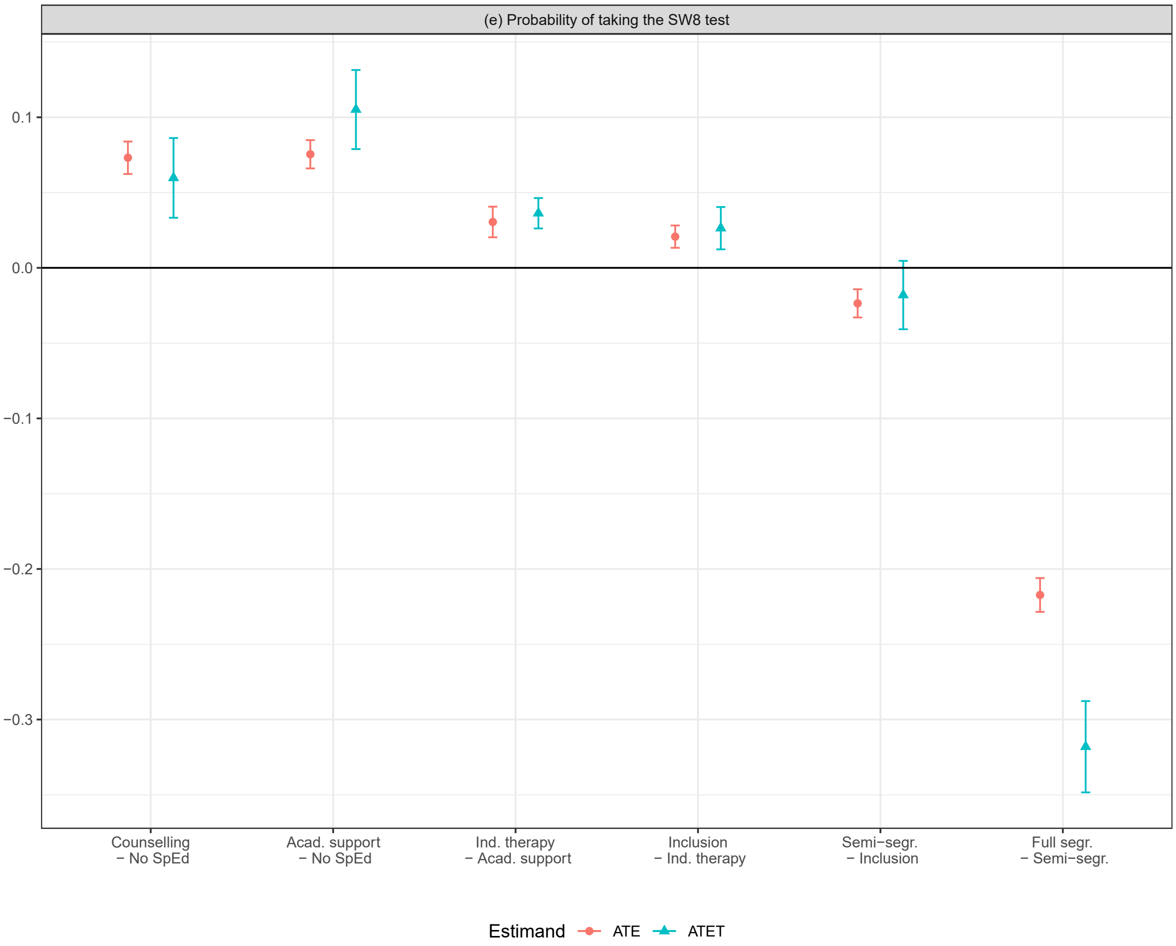
Figure A.1: Overlap of school characteristics for schools with inclusion and for schools with semi-segregation

*Notes:* This figure represents the distribution of schools with inclusion and of schools with semi-segregation along main school characteristics.



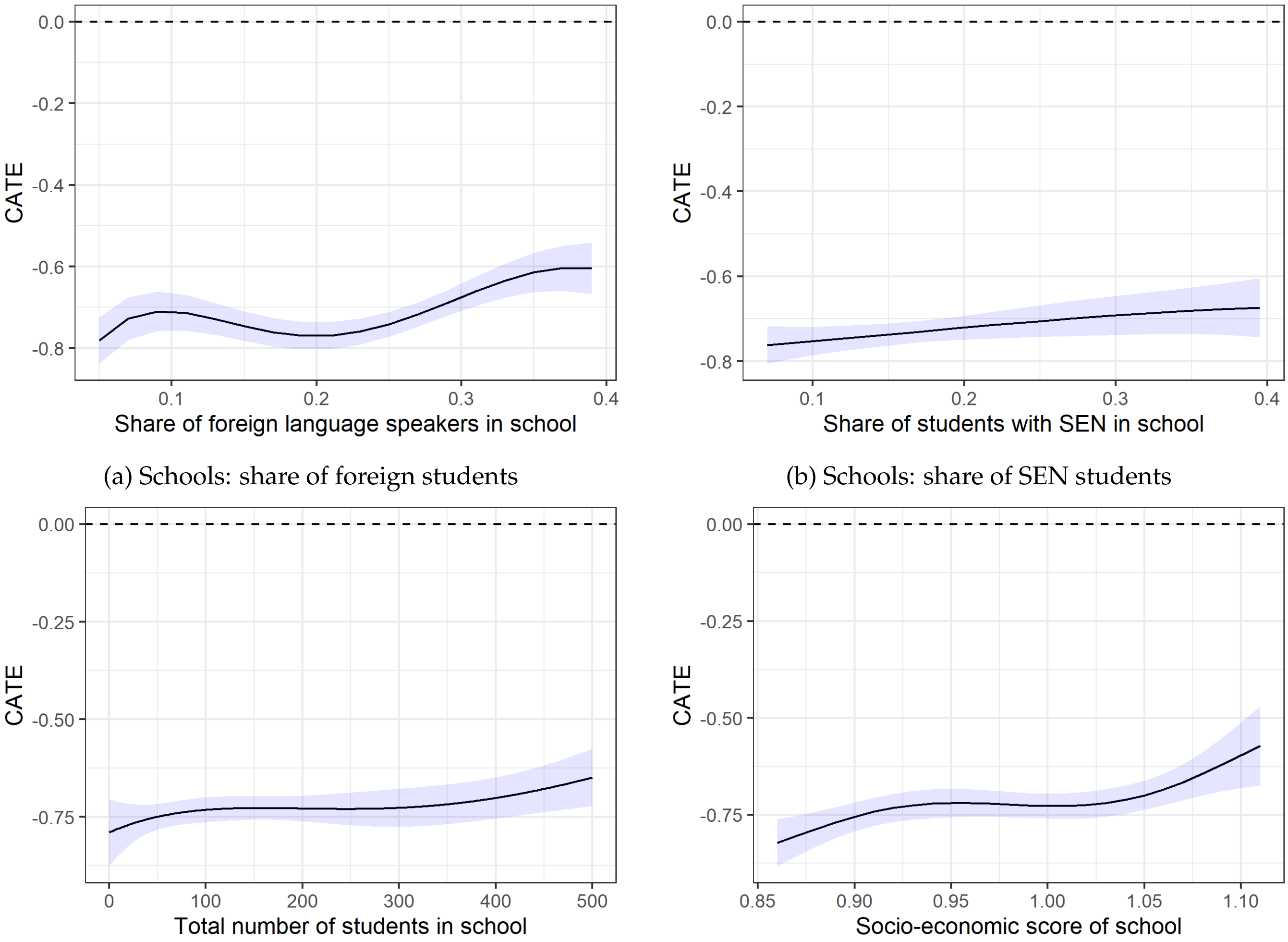
*Notes:* This figure shows the distribution of age and grade at registration to the School Psychological Service. The "Preschool" category groups three years of Kindergarten, which explains why it appears to be the largest category. The number of children who have been registered at the SPS (assigned and non-assigned to a particular therapy) is represented. *Source: SPS*.

Figure A.2: Distribution of age and grade at registration



*Notes:* This figure depicts relevant pairwise treatment effects for Special Education programs in St. Gallen. Each pairwise treatment effect is the effect of being assigned to the first program in comparison to the second program on one of the four outcomes presented in the panel headers. Both the treatment effect on the whole population (ATE) and on the population of the treated (ATET) are presented. “Ind. therapy” is the abbreviation for individual therapies, “Acad. support” for academic support, and “no SpEd” for receiving no program, “Semi-segr.” for semi-segregation (segregation in small classes), and “Full segr.” for full segregation (in special schools). Nuisance parameters are estimated using an ensemble learner that includes text representations presented in the “data” section. 95% confidence intervals are represented and are based on one sample $t - test$ for the ATE and the ATET. *Source: SPS*.

Figure A.3: Pairwise treatment effects for probability of taking the SW8 test



(a) Schools: share of foreign students

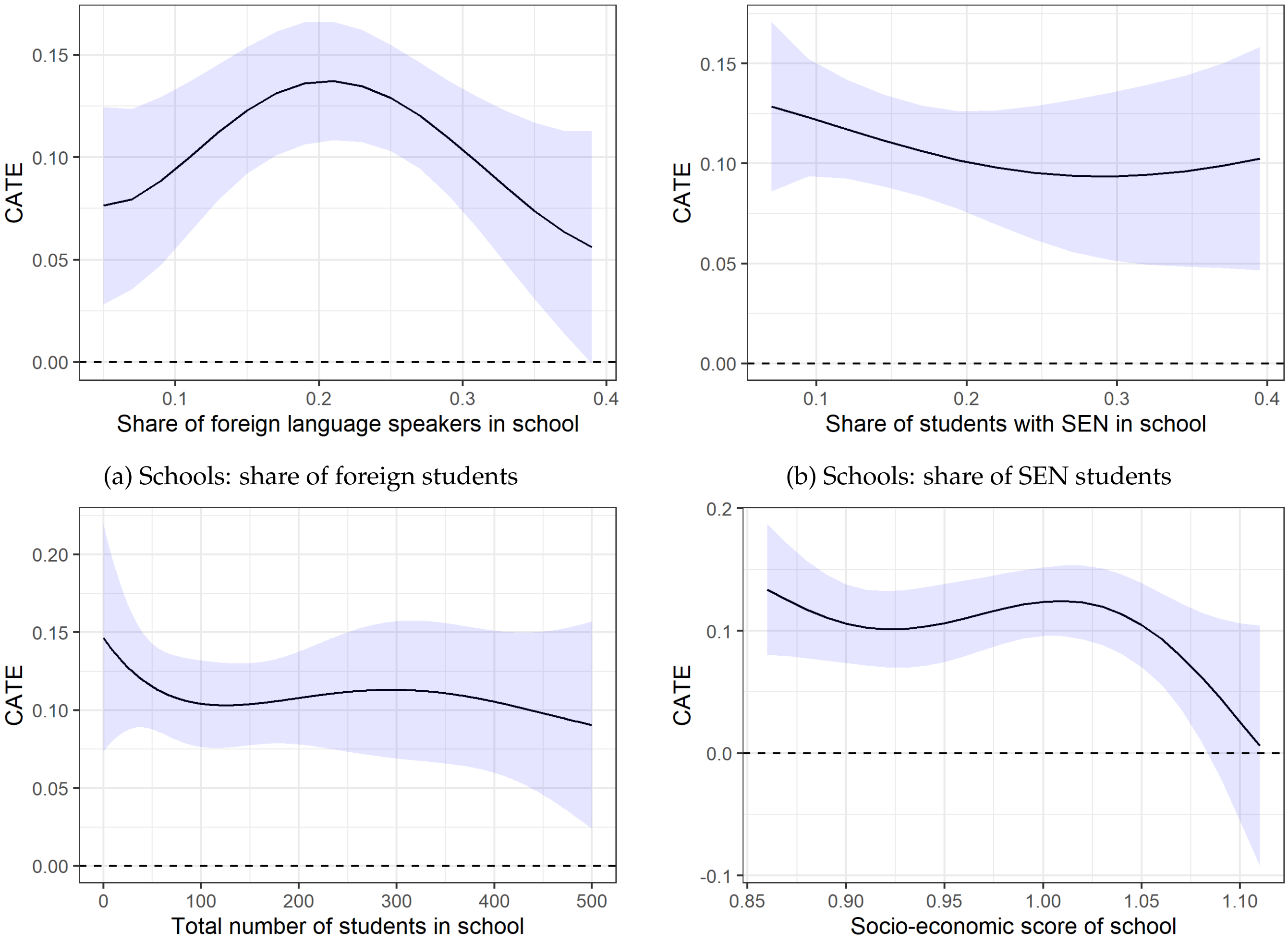
(b) Schools: share of SEN students

(c) Schools: total number of students

(d) Schools: "SES score"

*Notes:* Continuous Conditional Average Treatment Effects in academic performance for inclusion vs. semi-segregation are depicted. The CATEs show how much the ATE would change at different levels of school characteristics (on the $x$ axis).

Figure A.4: Academic performance: continuous CATE for school characteristics



(a) Schools: share of foreign students

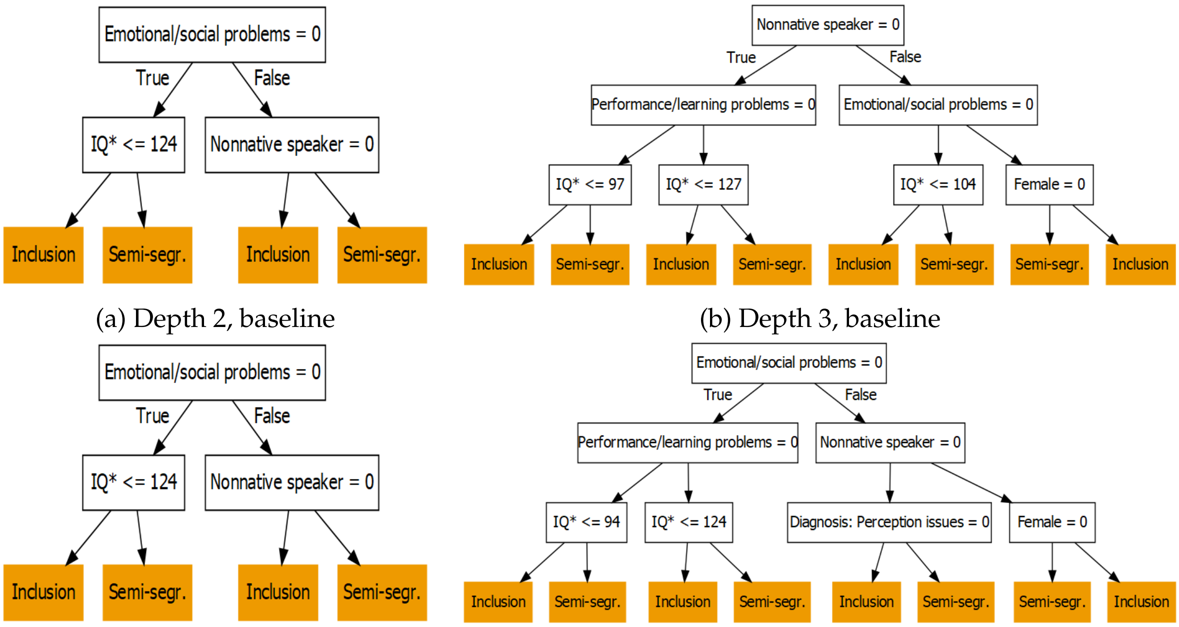
(b) Schools: share of SEN students

(c) Schools: total number of students

(d) Schools: "SES score"

*Notes:* Continuous Conditional Average Treatment Effects in the probability to be unemployed for inclusion vs. semi-segregation are depicted. The CATEs show how much the ATE would change at different levels of school characteristics (on the *x* axis).

Figure A.6: Probability of unemployment: continuous CATE for school characteristics



(a) Depth 2, baseline

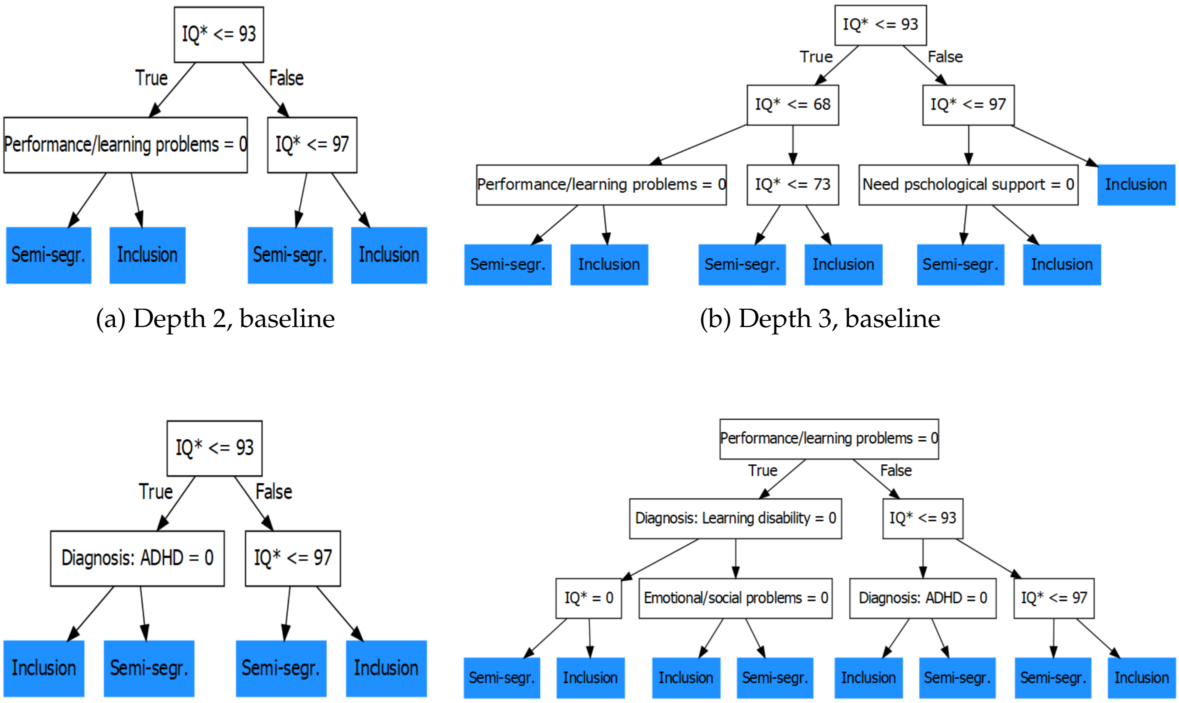
(b) Depth 3, baseline

(c) Depth 2, baseline + diagnosis

(d) Depth 3, baseline + diagnosis

*Notes:* Decision trees for optimal policy allocations of depth 2 and 3 are depicted. "Baseline" means that only baseline students' characteristics are included, and "baseline + diagnosis" includes both baseline characteristics and diagnosis characteristics extracted from psychological records with the dictionary approach. The IQ variable IQ* is the interaction between the IQ score and the indicator whether an IQ score has been taken.

Figure A.8: Optimal policy trees for test scores



(a) Depth 2, baseline

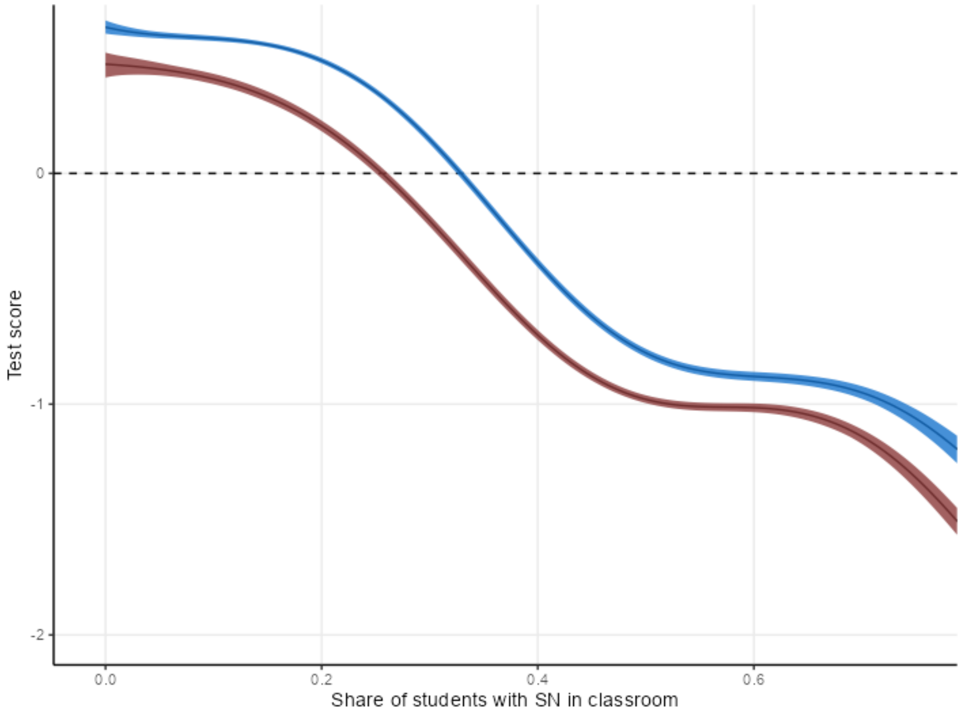
(b) Depth 3, baseline

(c) Depth 2, baseline + diagnosis

(d) Depth 3, baseline + diagnosis

*Notes:* Decision trees for optimal policy allocations of depth 2 and 3 are depicted. "Baseline" means that only baseline students' characteristics are included, and "baseline + diagnosis" includes both baseline characteristics and diagnosis characteristics extracted from psychological records with the dictionary approach. The IQ variable is the interaction between the IQ score and the indicator whether an IQ score has been taken.

Figure A.10: Optimal policy trees for probability to be employed



*Notes:* This graph depicts the spillover functions for the effect of the presence of students with SEN on the test scores of their peers with and without SEN in inclusive classrooms. All effects follow the identification strategy and results by Balestra, Eugster, and Liebert (forthcoming). Flexible spillover functions are estimated with an ensemble learner similar to the estimation procedure used in this paper. Clustered cross-validation procedures at the classroom level are implemented. 95% confidence intervals are represented.

Figure A.12: Classroom spillover effects of students with SEN on their peers without SEN.

# B | Appendix: Chapter 3

| | | Observations |
|---|---|---|
| Raw data | | 33,657 |
| Segregated special schools | – | 706 |
| Missing/implausible covariates | – | 15 |
| Missing/implausible test scores | – | 249 |
| Missing/implausible class size | – | 922 |
| Final sample | | 31,765 |
| Gifted students | – | 578 |
| Estimation sample | | 31,187 |

*Notes:* Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Table A.1: Construction of the sample

| | (1)<br>Missing Math<br>test score | (2)<br>Missing language<br>test score | (3)<br>Missing Math and<br>language test scores | (4)<br>Missing post-compulsory<br>education information | (5)<br>Missing occupation<br>profile information |
|---|---|---|---|---|---|
| Exposure to gifted classmates | 0.0004 | -0.0024 | -0.0010 | 0.0059 | -0.0075 |
| | (0.0029) | (0.0023) | (0.0017) | (0.0055) | (0.0056) |
| Mean outcome | 0.0050 | 0.0046 | 0.0027 | 0.2668 | 0.0816 |
| School-by-year FE | Yes | Yes | Yes | Yes | Yes |
| Observations | 31,406 | 31,406 | 31,406 | 31,406 | 15,552 |

*Notes:* *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$. Standard errors, shown in parentheses, are clustered at the classroom level. Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Table A.2: Attrition analysis

| | (1)<br>Composite<br>test score | (2)<br>Math<br>test score | (3)<br>Language<br>test score | (4)<br>Composite<br>test score | (5)<br>Math<br>test score | (6)<br>Language<br>test score |
|---|---|---|---|---|---|---|
| Proportion of gifted classmates | 1.141*** | 0.975*** | 1.000*** | 1.368*** | 1.271*** | 1.093*** |
| | (0.243) | (0.242) | (0.228) | (0.291) | (0.292) | (0.285) |
| | [0.057] | [0.049] | [0.050] | [0.068] | [0.064] | [0.055] |
| Female | -0.181*** | -0.354*** | 0.046*** | -0.173*** | -0.343*** | 0.049*** |
| | (0.015) | (0.014) | (0.014) | (0.016) | (0.016) | (0.016) |
| Proportion * Female | | | | -0.444 | -0.580* | -0.183 |
| | | | | (0.315) | (0.330) | (0.319) |
| Individual characteristics | Yes | Yes | Yes | Yes | Yes | Yes |
| Classroom characteristics | Yes | Yes | Yes | Yes | Yes | Yes |
| School-by-year FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 31,187 | 31,187 | 31,187 | 31,187 | 31,187 | 31,187 |

*Notes:* *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$. Standard errors, shown in parentheses, are clustered at the classroom level. The marginal effect of adding one gifted peer to a class of 20 is shown in brackets, and is defined as the coefficient divided by 20. Individual characteristics include gender, native German speaker, and age at test. Classroom characteristics include class size, share of females, share of native German speakers, and average age at test. Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Table A.3: Sensitivity to the specification of the treatment variable

| | (1)<br>Composite<br>test score | (2)<br>Math<br>test score | (3)<br>Language<br>test score | (4)<br>Composite<br>test score | (5)<br>Math<br>test score | (6)<br>Language<br>test score |
|---|---|---|---|---|---|---|
| Exposure to gifted classmates | 0.088*** | 0.065*** | 0.088*** | 0.107*** | 0.090*** | 0.095*** |
| | (0.023) | (0.023) | (0.020) | (0.025) | (0.026) | (0.024) |
| Female | -0.200*** | -0.368*** | 0.028** | -0.190*** | -0.355*** | 0.032** |
| | (0.012) | (0.013) | (0.012) | (0.014) | (0.014) | (0.014) |
| Exposure * Female | | | | -0.037 | -0.049* | -0.014 |
| | | | | (0.026) | (0.027) | (0.026) |
| Individual characteristics | Yes | Yes | Yes | Yes | Yes | Yes |
| Classroom characteristics | Yes | Yes | Yes | Yes | Yes | Yes |
| Teacher FE | Yes | Yes | Yes | Yes | Yes | Yes |
| School FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Year FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 31,187 | 31,187 | 31,187 | 31,187 | 31,187 | 31,187 |

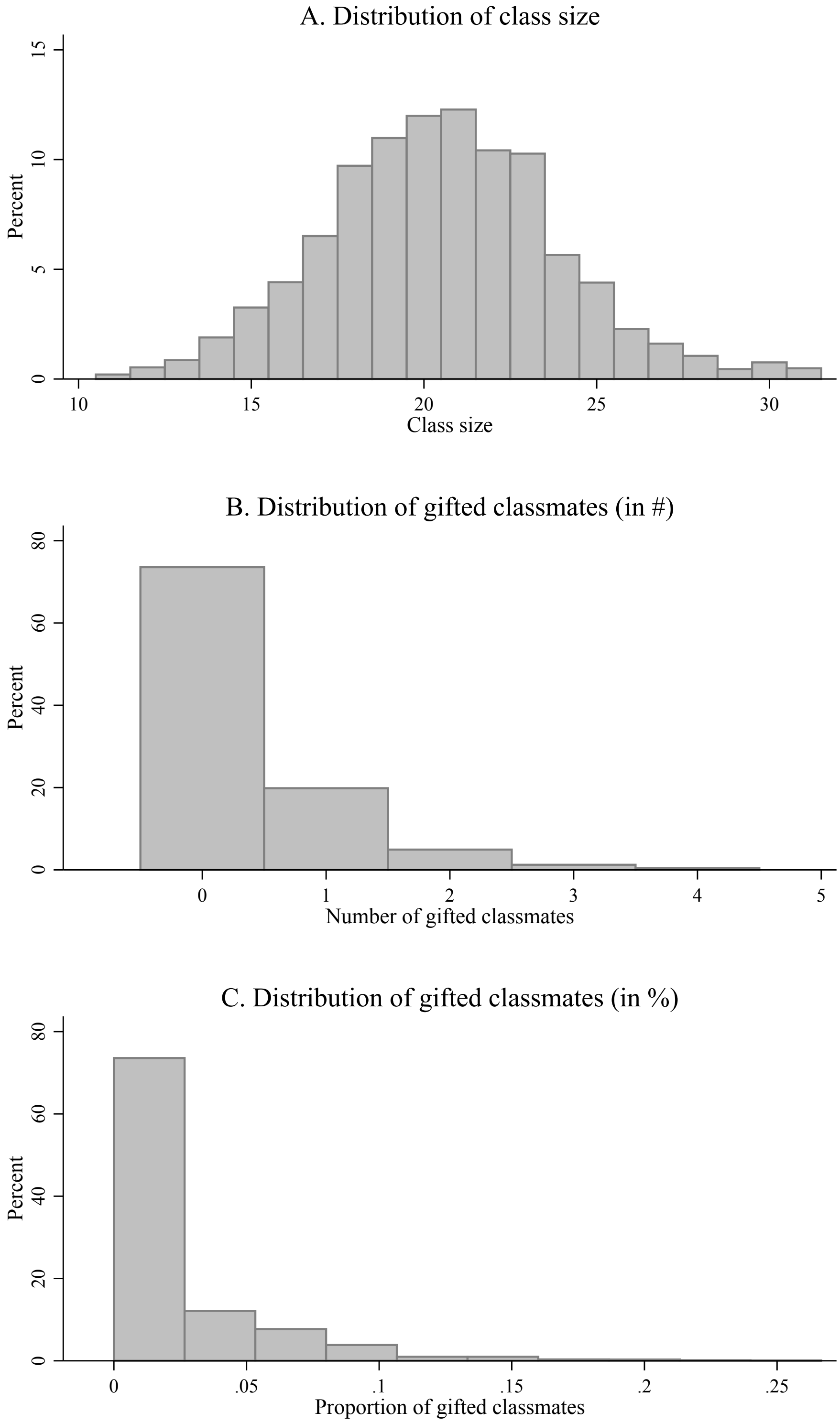
*Notes:* *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$. Standard errors, shown in parentheses, are clustered at the classroom level. Individual characteristics include gender, native German speaker, and age at test. Classroom characteristics include class size, share of females, share of native German speakers, and average age at test. Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Table A.4: Within-teacher identification

| | (1)<br>Composite<br>test score | (2)<br>Composite<br>test score | (3)<br>Composite<br>test score | (4)<br>Composite<br>test score |
|---|---|---|---|---|
| Exposure to gifted classmates | 0.081*** | 0.078* | 0.053* | 0.086*** |
| | (0.024) | (0.046) | (0.030) | (0.026) |
| Relative young | -0.133*** | | | |
| | (0.023) | | | |
| (Exposure to gifted classmates)* | 0.038 | | | |
| (Relative young) | (0.026) | | | |
| Native speaker | | 0.401*** | | |
| | | (0.025) | | |
| (Exposure to gifted classmates)* | | 0.019 | | |
| (Native speaker) | | (0.043) | | |
| Small class | | | -0.019 | |
| | | | (0.048) | |
| (Exposure to gifted classmates)* | | | 0.112** | |
| (Small class) | | | (0.047) | |
| Student-teacher same gender | | | | -0.009 |
| | | | | (0.014) |
| (Exposure to gifted classmates)* | | | | 0.018 |
| (Student-teacher same gender) | | | | (0.027) |
| Individual characteristics | Yes | Yes | Yes | Yes |
| Classroom characteristics | Yes | Yes | Yes | Yes |
| School-by-year FE | Yes | Yes | Yes | Yes |
| Observations | 31,187 | 31,187 | 31,187 | 31,187 |

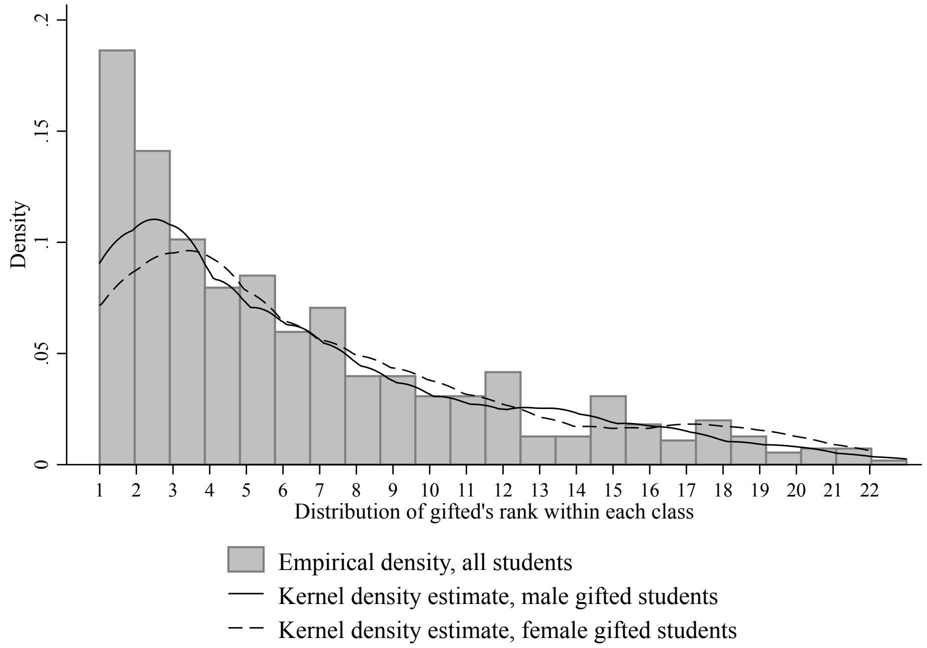
*Notes:* *** $p < 0.01$, ** $p < 0.05$, and * $p < 0.10$. Standard errors, shown in parentheses, are clustered at the classroom level. Individual characteristics include gender, native German speaker, and age at test. Classroom characteristics include class size, share of females, share of native German speakers, and average age at test. Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Table A.5: Heterogeneity analysis: relative age, native speaker, class size, and teacher's gender



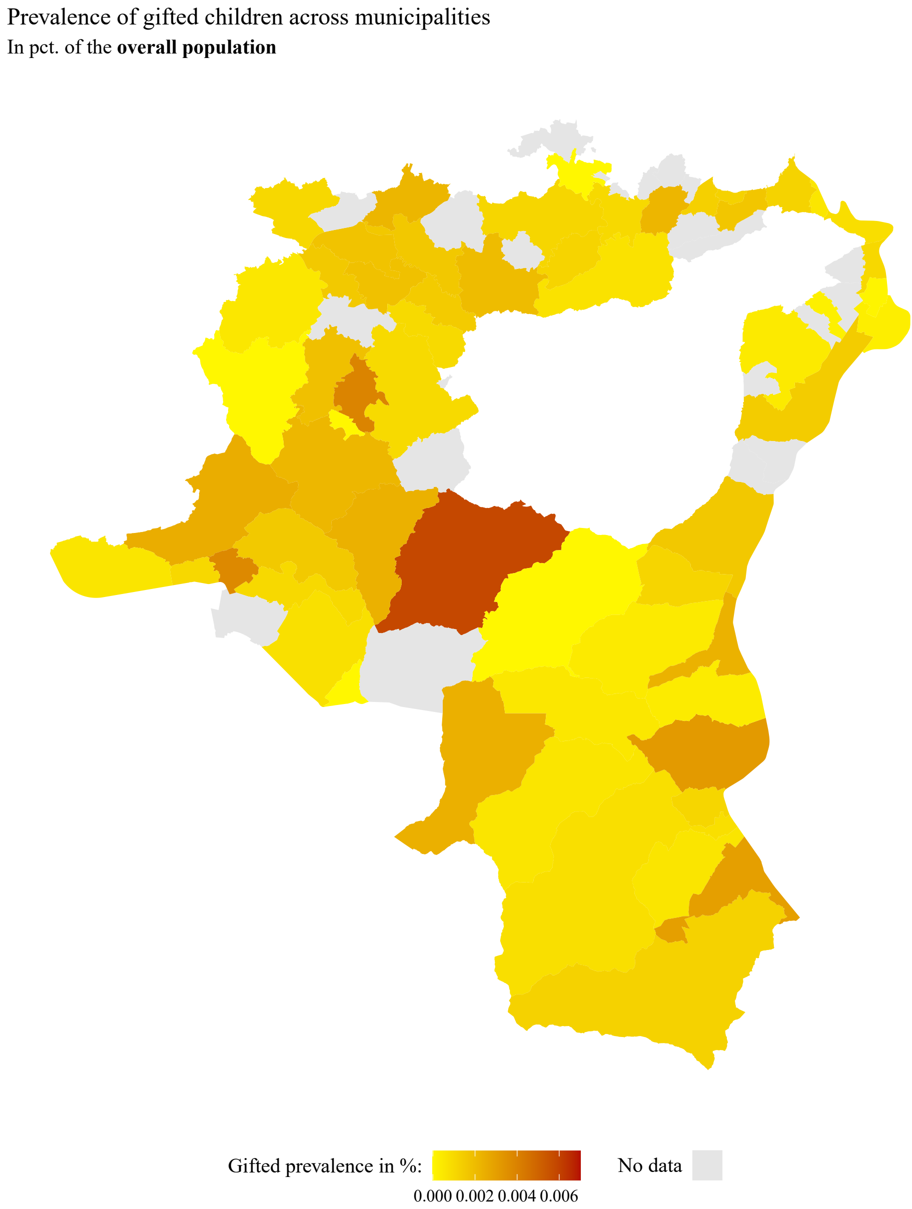
*Notes:* Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Figure A.1: Distribution of class size and gifted classmates



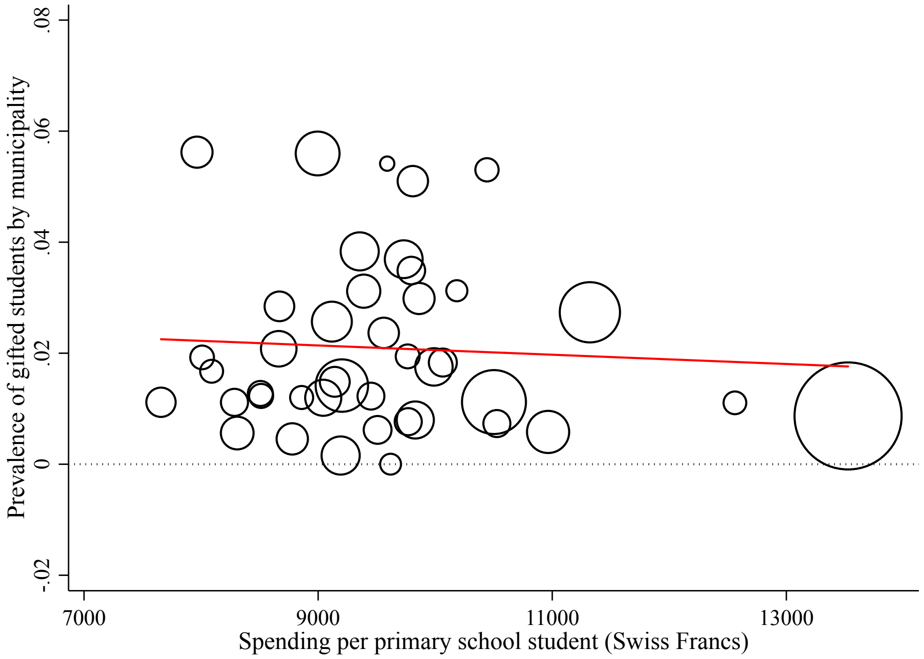
*Notes:* Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Figure A.2: Distribution of outcome ranks of gifted children in their classroom


Prevalence of gifted children across municipalities
In pct. of the **overall population**
Gifted prevalence in %:
0.000 0.002 0.004 0.006
No data

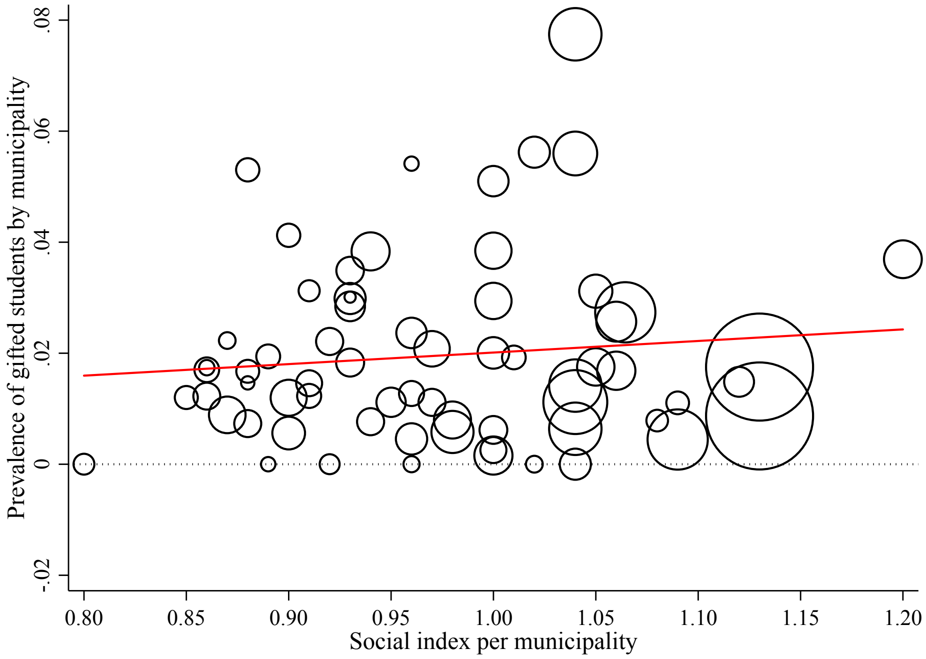
*Notes:* Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Figure A.3: Prevalence of gifted students by municipality



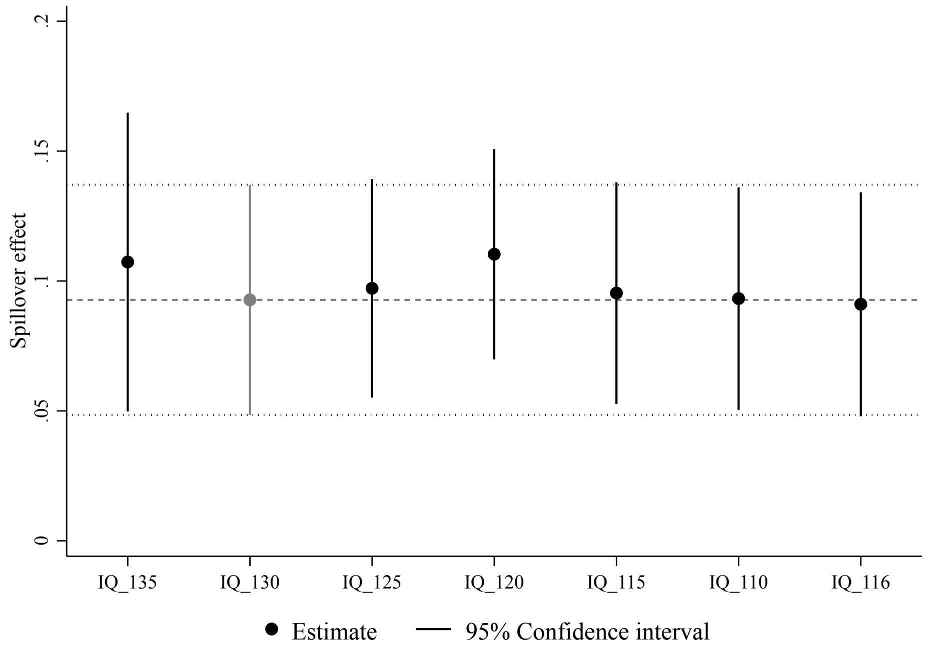
*Notes:* Each circle represents one municipality, the dimension of the circle represents the size of the municipality, and the red solid line is a linear fit that summarizes the relation between exposure and spending. Data refer to 2017 and are from the official accounts published by each municipality at the end of the fiscal year.

Figure A.4: Per-student spending by exposure to gifted classmates



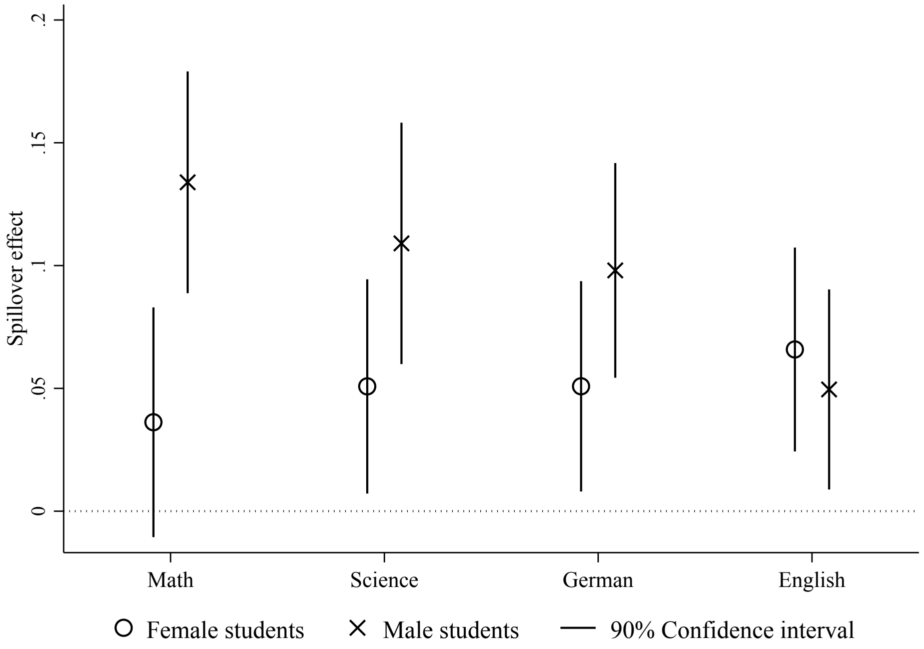
*Notes:* each circle represents one municipality, the dimension of the circle represents the size of the municipality, and the red solid line is a linear fit that summarizes the relation between exposure and index. Data refer to 2007 (at baseline) and are provided by the Competence Center for Statistics within the Department of Economic Affairs of the Canton of St. Gallen.

Figure A.5: Social index by exposure to gifted classmates



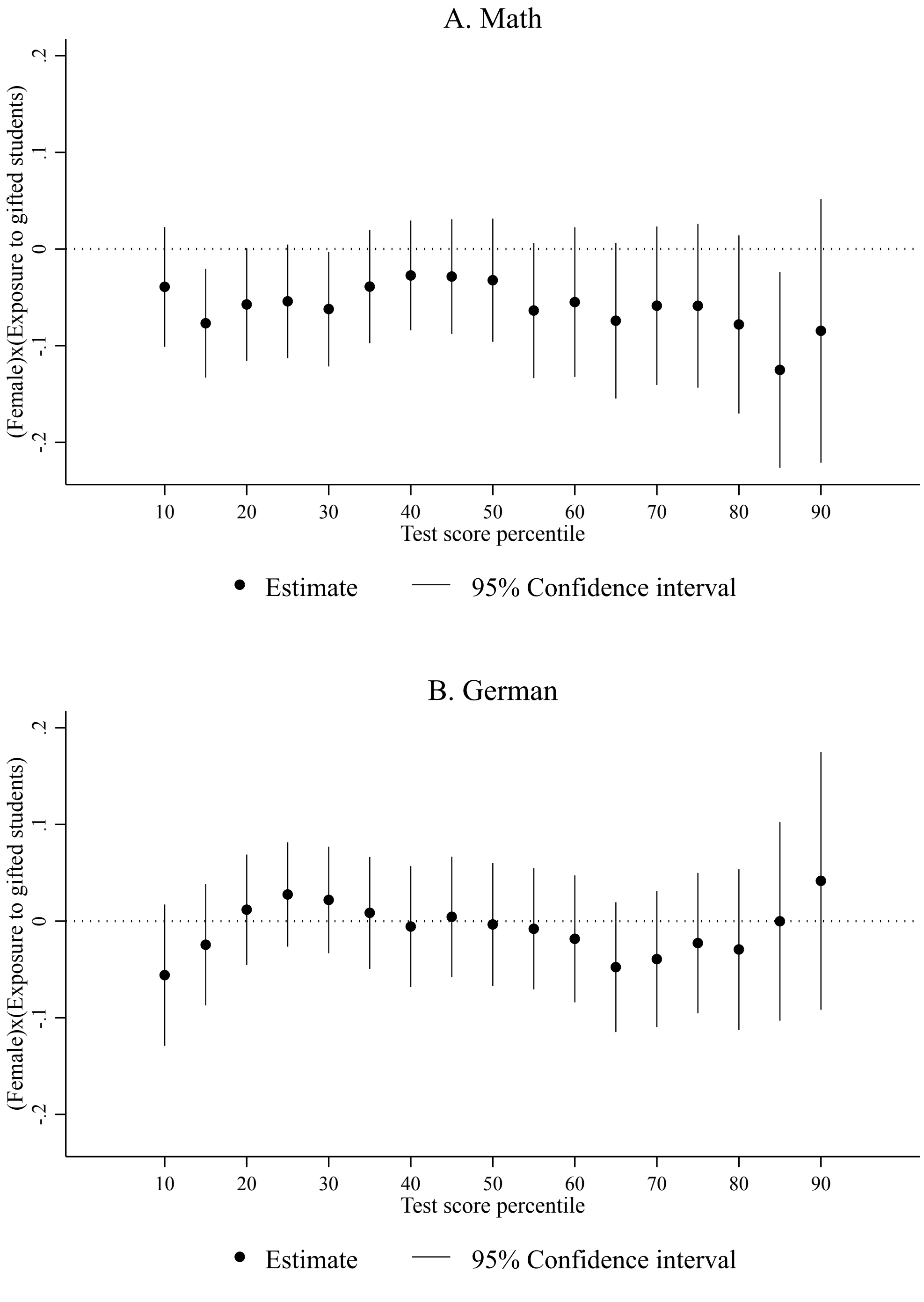
*Notes:* Results are based on separate regressions on the same estimation sample, which comprises students with IQ below 110 (29,862 observations). The gray estimate is based on the the main IQ threshold of the paper. The last estimate (IQ-116) uses the alternative IQ threshold for non-native speakers of 116 points. Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Figure A.6: Sensitivity to the IQ threshold for classifying a student as gifted



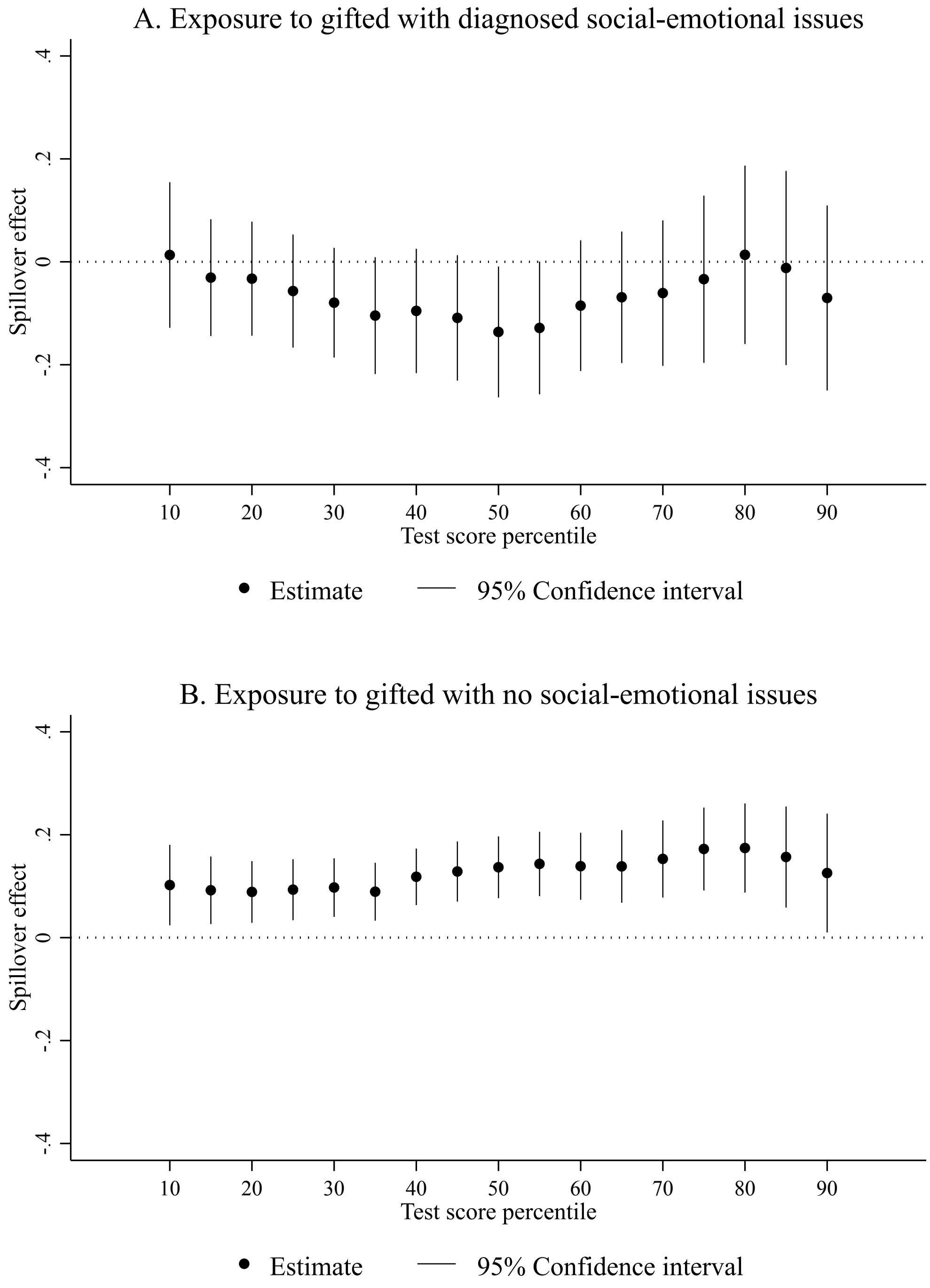
*Notes:* Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Figure A.7: Spillovers by school subjects



*Notes:* Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Figure A.8: Quantile treatment effect of exposure to gifted classmates for female students and male students

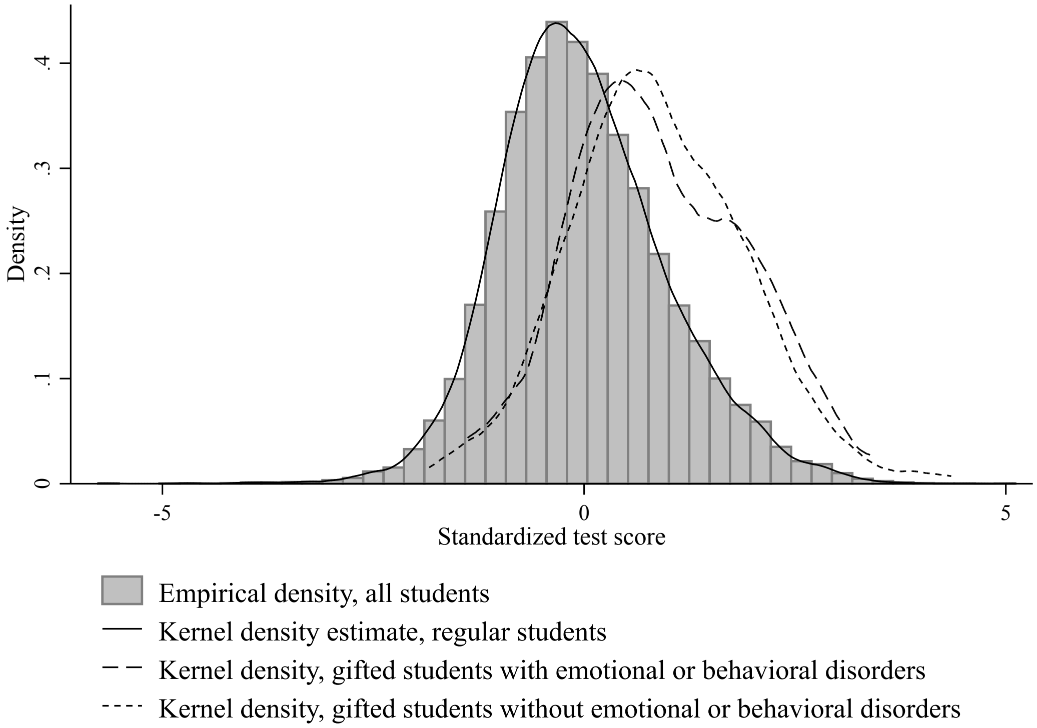


A. Exposure to gifted with diagnosed social-emotional issues

B. Exposure to gifted with no social-emotional issues

*Notes:* Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Figure A.9: Quantile treatment effect of exposure to gifted students with and without other concurring diagnosis

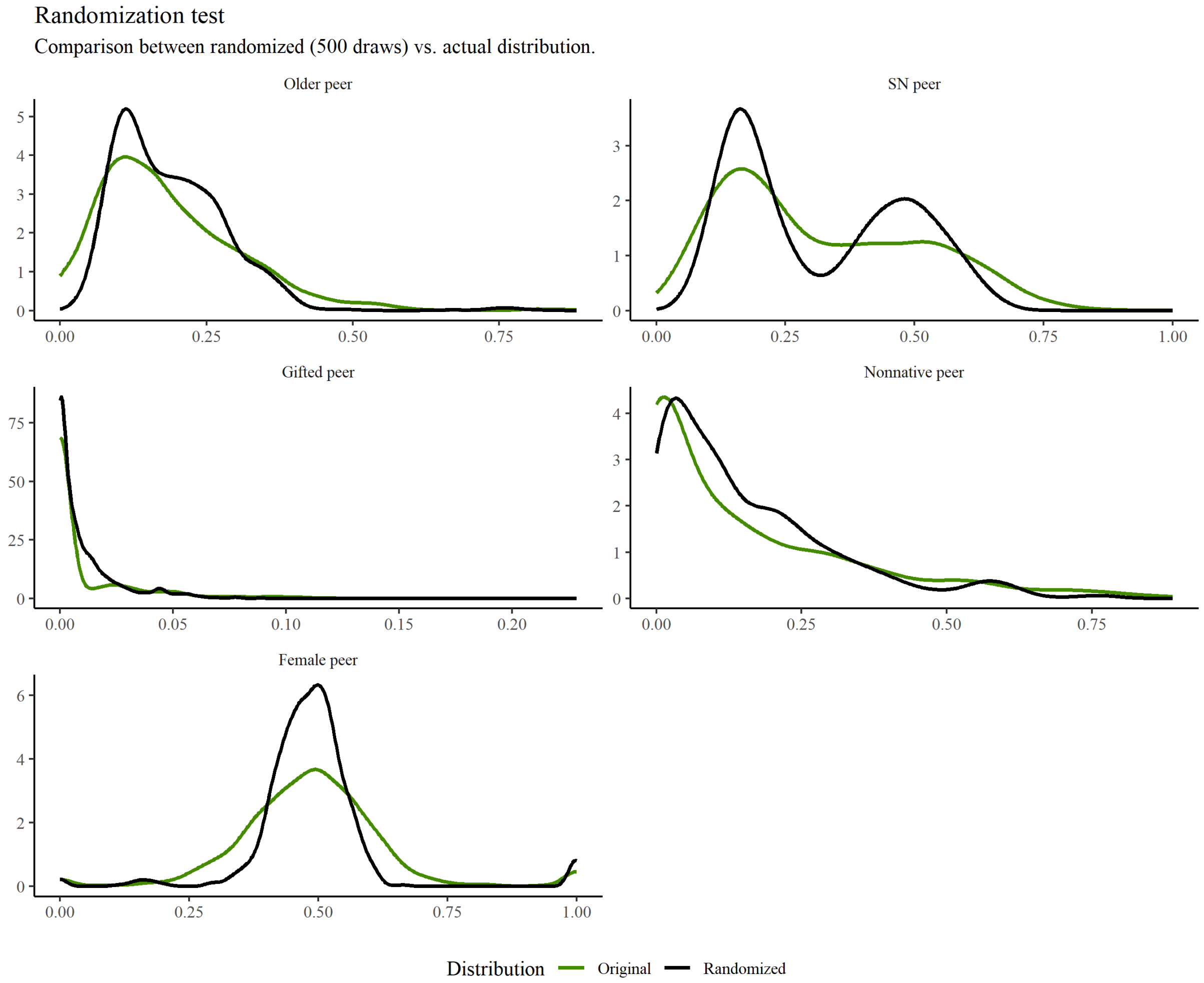


*Notes:* Data are from the School Psychological Service St. Gallen and the Stellwerk test service provider.

Figure A.10: Distribution of test scores for gifted children with and without emotional or behavioral disorders

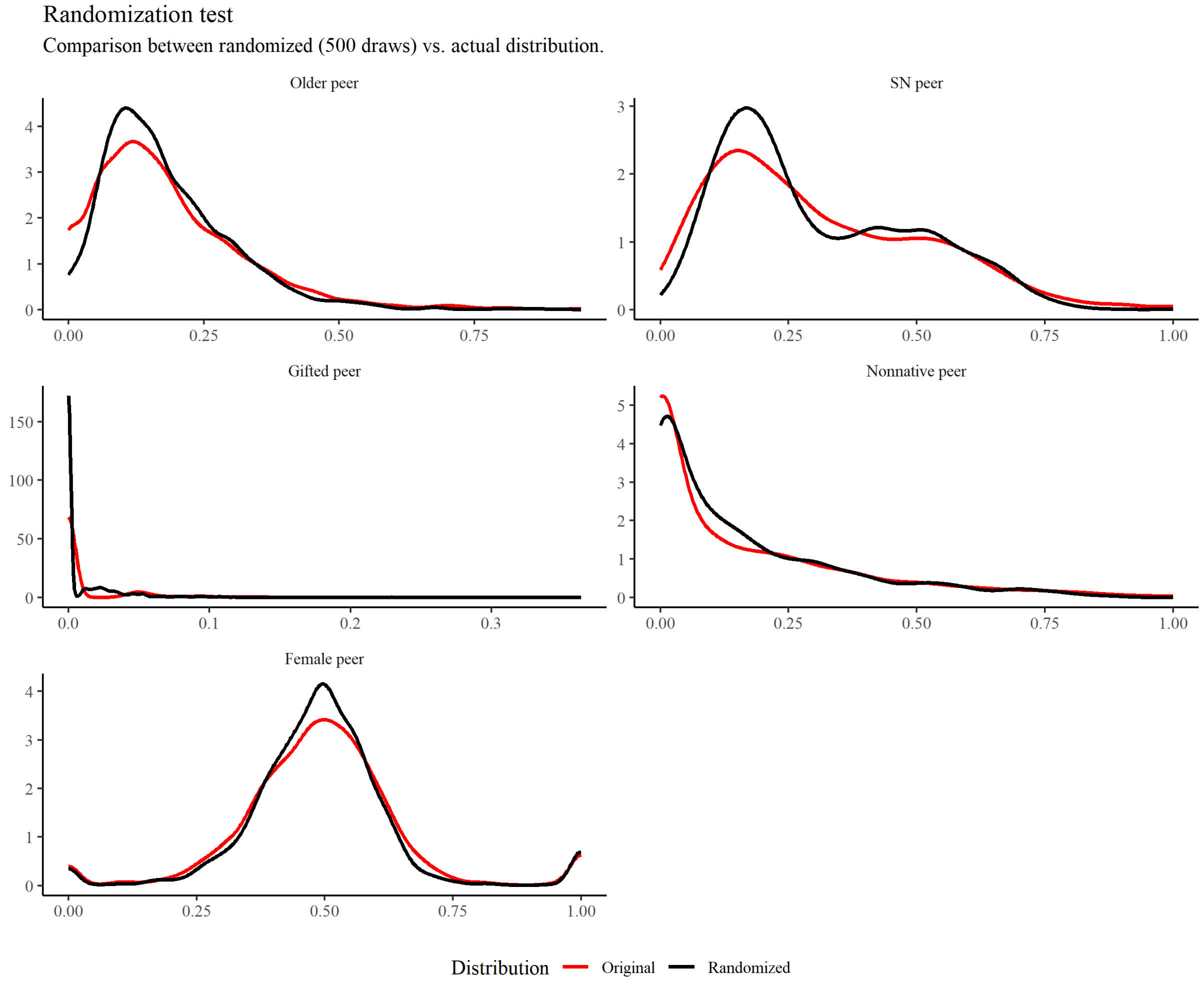
# C | Appendix: Chapter 4

## C.1 Appendix: Supplementary Material



*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.1: Balancing check for cohort identification



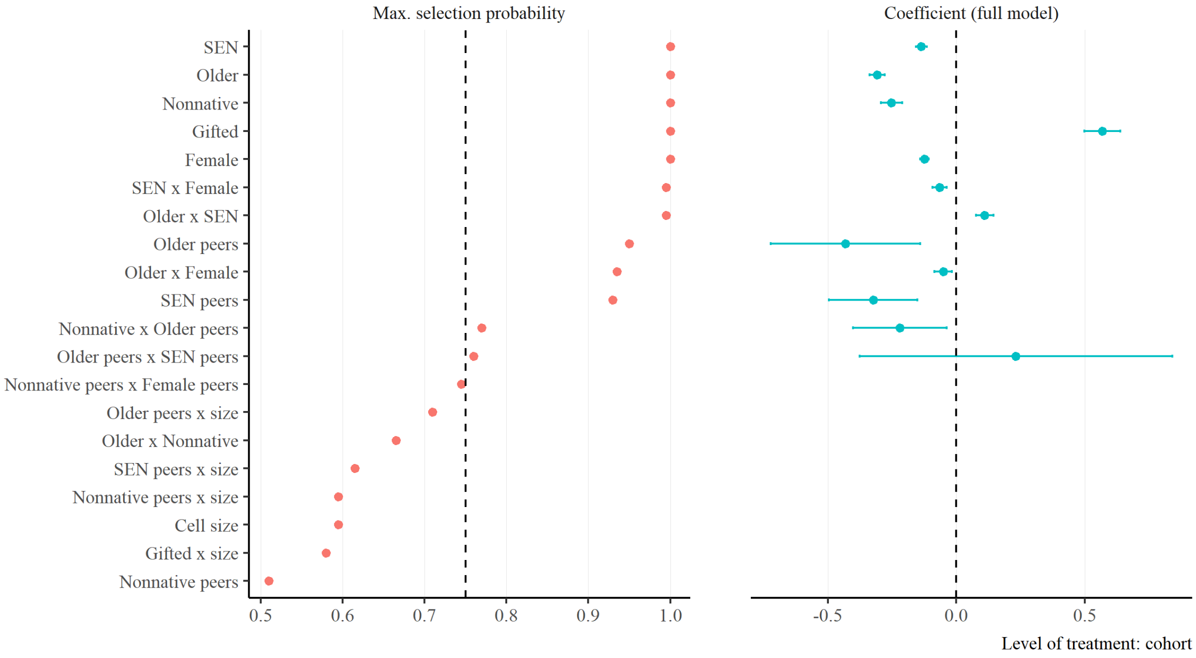
*Notes:* This graph shows the actual and the randomized (simulated) distribution of the proportion of peers of a given type within school-track-years. An observation is a classroom. We randomly draw observations at the school-track-year level 500 times with replacement and compare the randomly sampled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

Figure A.2: Balancing check for classroom identification

## C.2 Appendix: Stable selection with interactions of degree 2

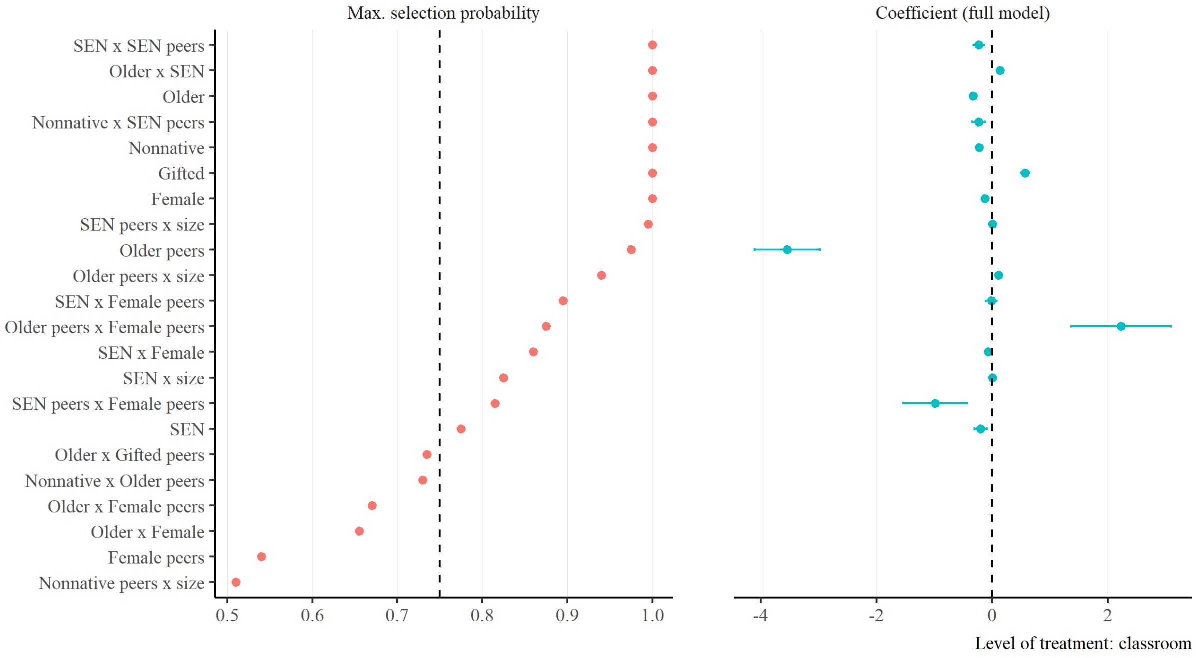
As presented in Figures B.1 and B.2, 13 variables are selected at the cohort level, and 19 at the classroom level. Among the variables selected at the cohort level, only four variables are peer effects: the effect of older peers, and the effect of peers with SEN. These two peer effects are heterogeneous: 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.



*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 × 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 4.3.3. Effect sizes and confidence intervals are presented for information only and must be interpreted with caution.

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 × 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 4.3.3. 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

## C.3 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 A.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.

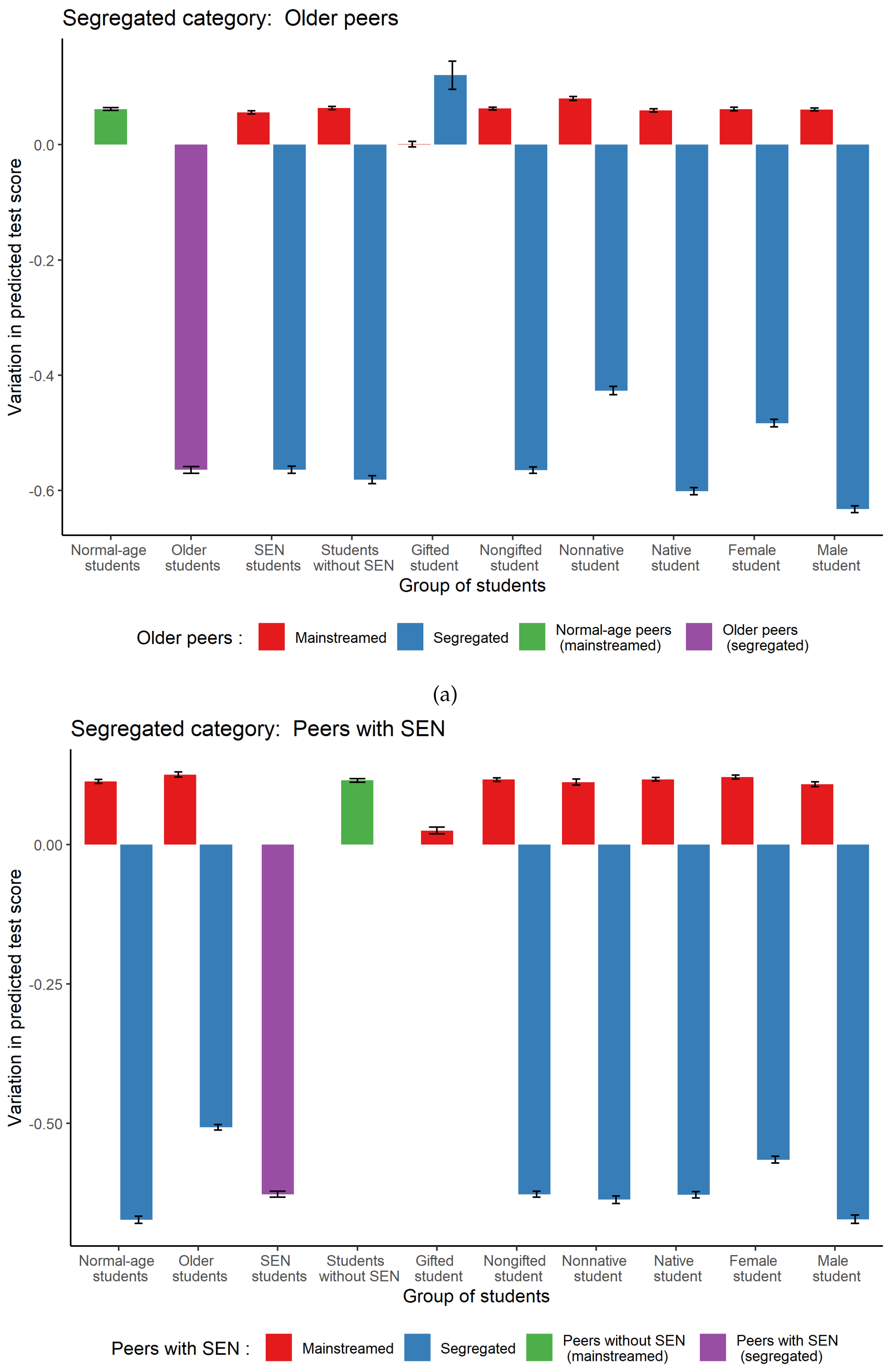
**Results for the *CACE*** We look at the variations in predicted aggregate test scores for each type of student. Figure B.3 and Figure B.5 present the gains and losses for students of all categories when older students (Figure B.3a.), when students with SEN (Figure B.3b.), when nonnative students (Figure B.5a.), and when female students (Figure B.5b.) are segregated. These are group test score averages under the different counterfactual regimes.

| | Randomized regime | Segregated regime | Difference |
|---|---|---|---|
| | $\frac{1}{N}\sum_{i=1}^{N}\left[\hat{Y}_{ic}^{\text{random}}\right]$ | $\frac{1}{N}\sum_{i=1}^{N}\left[\hat{Y}_{ic}^{\text{segr.}}\right]$ | *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 A.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.

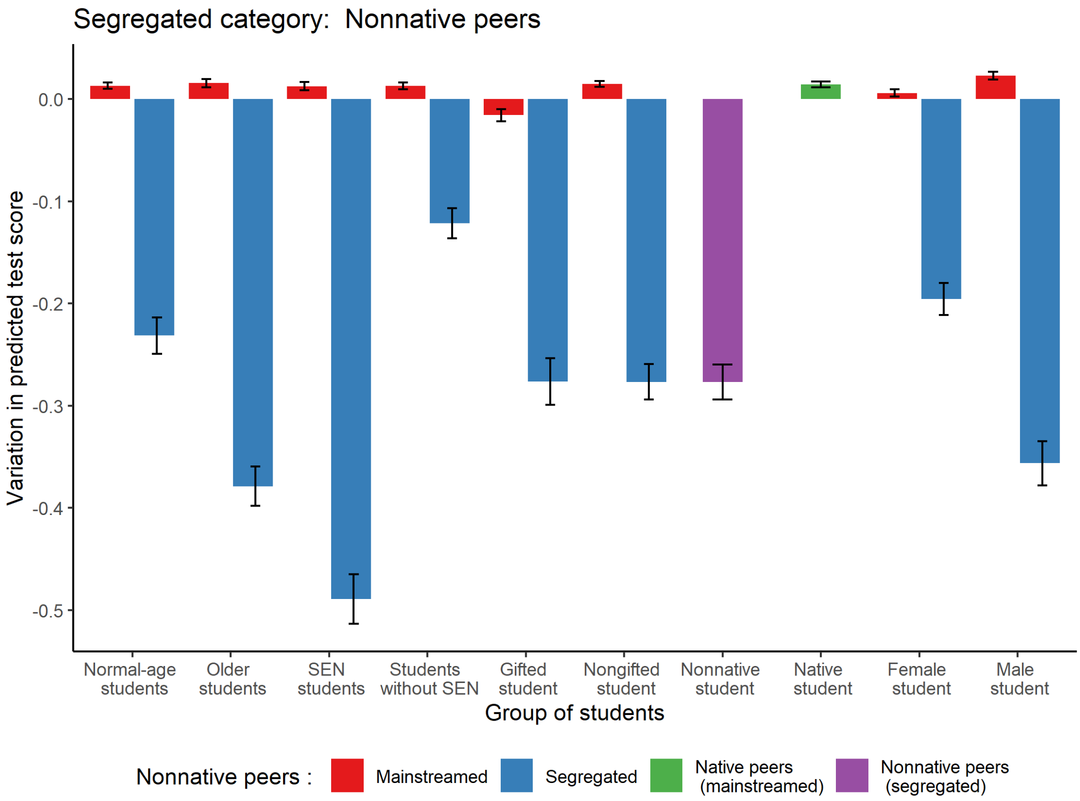


(a)

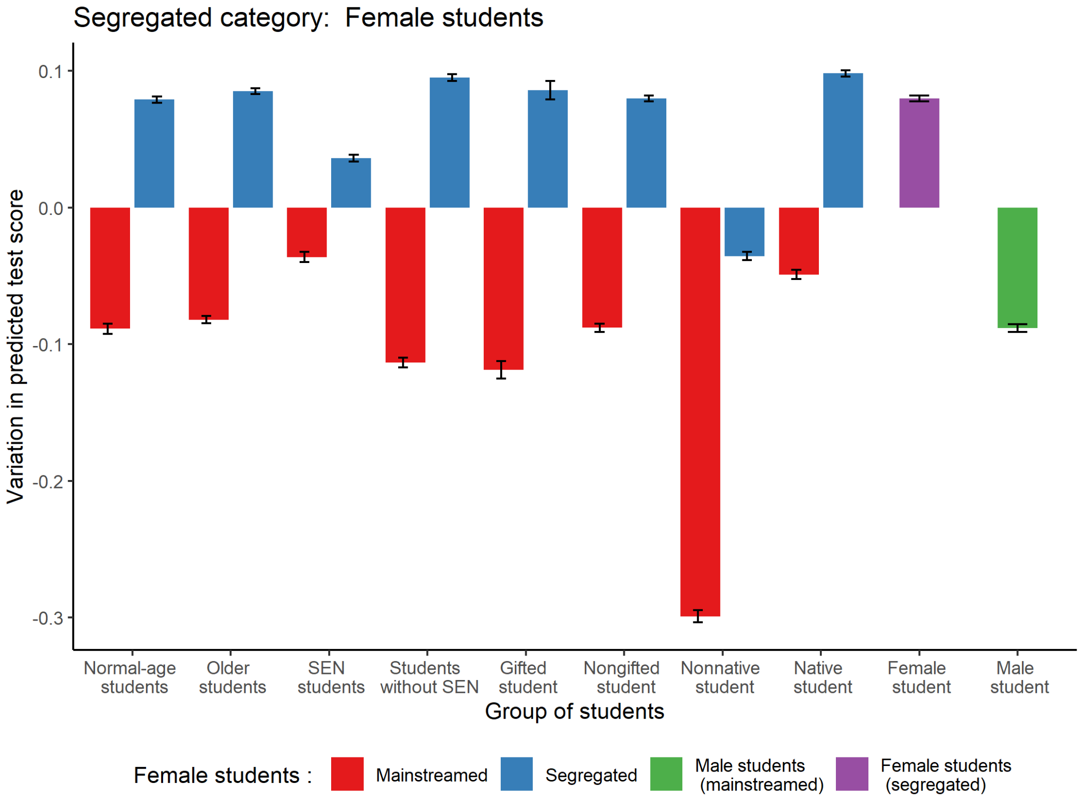
(b)

*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).

Figure B.3: CACE: older peers and peers with SEN are segregated



(a)



(b)

*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).

Figure B.5: CACE: nonnative and female students are segregated

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 B.3a, 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., Pregaldini, Backes-Gellner, and Eisenkopf, 2020; Eisenkopf et al., 2015). 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.

# Curriculum Vitae

## Aurélien Sallin

Born May 28, 1988, from Villaz-Saint-Pierre, Switzerland.

### Education

| | |
|---|---|
| 2018-2022 | PhD in Economics and Finance<br>University of St. Gallen, Switzerland |
| 2014-2018 | Master of Arts in Philosophy<br>University of Fribourg, Switzerland |
| 2015 | Exchange Semester<br>University of St. Gallen, Switzerland |
| 2014-2017 | Master of Arts in Economics<br>University of Fribourg, Switzerland |
| 2011-2014 | Bachelor of Arts in Philosophy & Economics<br>University of Fribourg, Switzerland |

### Employment

| | |
|---|---|
| 2017-2022 | Research Assistant and Teaching Assistant<br>University of St. Gallen |
| 2020-present | Logic-Based-Therapy Practitioner<br>National Philosophical Counseling Association (NPCA) |
| 2015-present | Strategic advisor<br>InterGifted |
| 2015-2017 | Research Assistant<br>University of Fribourg, Department of Economics |
| 2014 | Research Assistant<br>University of Fribourg, Department of Philosophy |
| 2009-2011 | Pontifical Swiss Guard<br>Vatican City |