Universität St.Gallen

Estimating returns to special education: combining machine learning and text analysis to address confounding

Aurélien Sallin

October 2021 Discussion Paper no. 2021-09

School of Economics and Political Science,
Department of Economics

University of St.Gallen

Editor: Mirela Keuschnigg
University of St.Gallen
School of Economics and Political Science
Department of Economics
Müller-Friedberg-Strasse 6/8
CH-9000 St.Gallen
Email seps@unisg.ch

Publisher: School of Economics and Political Science
Department of Economics
University of St.Gallen
Müller-Friedberg-Strasse 6/8
CH-9000 St.Gallen

Electronic Publication: http://www.seps.unisg.ch

Estimating returns to special education: combining machine learning and text analysis to address confounding[1]

Aurélien Sallin

Author's address:
Aurélien Sallin
Swiss Institute for Empirical Economic Research SEW-HSG
Varnbüelstrasse 14
9000 St. Gallen
Email aurelien.sallin@unisg.ch
Website https://asallin.github.io

[1] I am grateful to my supervisor Beatrix Eugster, as well as Simone Balestra, Uschi Backes-Gellner, Caroline Chuard, Jennifer Harvey Sallin, Martin Huber, Michael Knaus, Edward Lazear, Michael Lechner, Helge Liebert, Fanny Puljic, participants of 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.

Abstract

While the number of students with identified special needs is increasing in developed countries, there is little evidence on academic outcomes and labor market integration returns to special education. I present results from the first ever study to examine short- and long-term returns to special education programs using recent methods in causal machine learning and computational text analysis. I find that special education programs in inclusive settings have positive returns on academic performance in math and language as well as on employment and wages. Moreover, I uncover a positive effect of inclusive special education programs in comparison to segregated programs. However, I find that segregation has benefits for some students: students with emotional or behavioral problems, and nonnative students. Finally, using shallow decision trees, I deliver optimal placement rules that increase overall returns for students with special needs and lower special education costs. These placement rules would reallocate most students with special needs from segregation to inclusion, which reinforces the conclusion that inclusion is beneficial to students with special needs.

Keywords

returns to education, special education, inclusion, segregation, causal machine learning, computational text analysis

JEL Classification

H52, I21, I26, J14, C31, Z13

1 Introduction

A growing number of students in OECD countries are identified with special needs (SEN)[1]. 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). Moreover, inclusion of students with special needs in mainstream education has been set as an educational objective by developed countries since the late 1990's.[2] To this end, most OECD countries have reduced segregation of SEN students and increased the involvement of mainstream public education by offering a variety of services such as alternative teaching methods and curricula, Individualized Education Programs (IEPS), and increased staff to accommodate individualized support within the main classroom.[3]

Despite the increasing number of students with SEN and the push to implement inclusive education, evidence on how special education (SpEd) placements - and inclusive placements in particular - affect academic and labor market returns of students with SEN is 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; Heckman et al., 2010), it is crucial to provide teachers, parents, psychologists and policy makers with insights on 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 necessity is particularly urgent in light of the considerable additional financial costs SpEd programs generate for public schools in comparison to standard education (Duncombe and Yinger, 2005). 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 times as much as educating a mainstreamed student.[4]

In this paper, I set out to investigate returns to special education on academic performance, labor participation, use of disability insurance, and wages. Moreover, I analyze returns to special education in a comprehensive way, and assess returns of six different SpEd programs. The first four programs are offered in inclusive academic settings, and comprise 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

[1] 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.

[2] 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".

[3] 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).

[4] For instance, 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).

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 in 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 special education in the Swiss State of St. Gallen.

I conduct these analyses in the context of the Swiss education system, an academic setting which offers ideal conditions for the investigation of returns to SE. A first ideal feature is that 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 heterogeneity in program availability and treatment assignment across schools, and that many schools implement both inclusive as well as segregated programs. A second feature is the fact that schools in St. Gallen were strongly encouraged to implement inclusive programs instead of segregated ones. However, not all schools started replacing segregated programs with inclusive programs at the same time. This strong institutional shift away from segregation means that students with similar special needs were sent either to inclusive or segregated settings depending on their schools. Third, special needs are assessed and determined by the school psychological service, an independent and centralized institution that provides children and their parents with a diagnosis and that assigns each student to a particular program in their own school. This ensures that treatment is given by professional psychologists, and not by parents, teachers, or schools.

Special education programs are difficult to evaluate because 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 psychological records and session transcripts written by psychologists and caseworkers about each SEN student. These records allow me to gain a deep understanding of the treatment assignment process, 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, natural language processing (NLP) and information retrieval in social sciences 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 (from most to least inclusive interventions), as well as students assigned to SpEd interventions with students that were referred but not treated. I find that, among all special education programs, inclusive programs

pay off: first, returns to SpEd programs provided in mainstream education are mostly positive or null in comparison to receiving no SpEd at all (but being referred to the school psychological service for assessment). I present evidence that targeted individual therapies (such as speech therapy, dyslexia therapies, etc.) are effective at remediating 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 15 percentage points more 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 very similar in all their observed characteristics. 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. 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 students, are the students who benefit the most from semi-segregation in comparison to inclusion. My empirical conclusions however do not extend to students with SEN who are fully segregated (in special schools), as these students differ substantially in their characteristics from other students. 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.

I further explore optimal policy allocations to inclusive and segregated SpEd programs and make placement recommendations to reach higher aggregate school performance and better labor market integration. I propose a set of optimal policies using shallow decision trees ("policy trees") (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 lesser extent, labor market integration at lower overall costs. Higher outcomes for students with SEN are reached by sending younger students with social and emotional problems or needs for additional psychological support to semi-segregation. I further conduct welfare computations to see whether mainstreaming students would harm students without SEN. I extend 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 not harm other students if included peers with SEN are distributed evenly across classrooms.

The present paper contributes to the understanding of returns to SpEd programs. Most studies use Special Education as a single, all-encompassing treatment intervention, and compare students in SpEd with students out of SE. 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.[5] Studies have shown moderate effectiveness of SpEd programs on the academic performance of SEN students (Schwartz, Hopkins, and Stiefel, 2021; Keslair, Maurin, and McNally, 2012; Harrison et al., 2013; Lavy and Schlosser, 2005)[6] and positive returns for

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

[6] 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

SEN students with learning and/or emotional disabilities (Hanushek, Kain, and Rivkin, 2002). In addition, evidence on long-term effects of SpEd placements is scant: while Ballis and Heath (forthcoming) find that students experience long-run benefits from SpEd placements, other studies report long-term negative impacts of SpEd.[7] As far as I know, this is the first study that assesses short- and long-term returns to SpEd programs at a granular level by ordering 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 special education 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 address 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 causal investigations of the effects of inclusion in comparison to segregation. On the one hand, existing research offers inconclusive results 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)[8] 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.

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).

[7] 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.

[8] 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 higher motivation, and 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).

2 Background and Data

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 in the main classroom of SEN students rather than semi-segregation. Inclusion was promoted as the main SpEd intervention tool (Wolter and Kull, 2006) and as a direct substitute for semi-segregation (Häfeli and Walther-Müller, 2005). As a result, the share of students sent to segregated schooling has decreased while the share of students sent to inclusive schooling has increased since the Equality Act. The share of segregated Swiss SEN students remains relatively high in international comparison (Sermier-Dessemontet, Benoit, and Bless, 2011; Wolter and Kull, 2014), and this share 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.[9]

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 measures and programs for SEN children.[10] These measures encompass counseling, academic support, individual therapies, inclusive measures, semi-segregated measures, and fully segregated measures. Counseling in this context 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. Academic support refers to tutoring for children needing additional support for their homework or for learning. Individual therapies refer to one-to-one or small group sessions that target particular learning disabilities (speech therapy, dyslexia or dyscalculia therapy, for instance). Individual therapies take place during class time. My inclusion measure refers to all students who received individual inclusive SpEd (ISF, "*Integrierte Schülerförderung*"). Students receiving inclusive SpEd are provided adapted and goal-oriented complementary teaching by a SpEd teacher who works in the main classroom alongside the main teacher. Semi-segregated measures refer 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. Finally, 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).

Figure 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 SEN students in 1998 to around 30% in 2010, whereas the

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

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

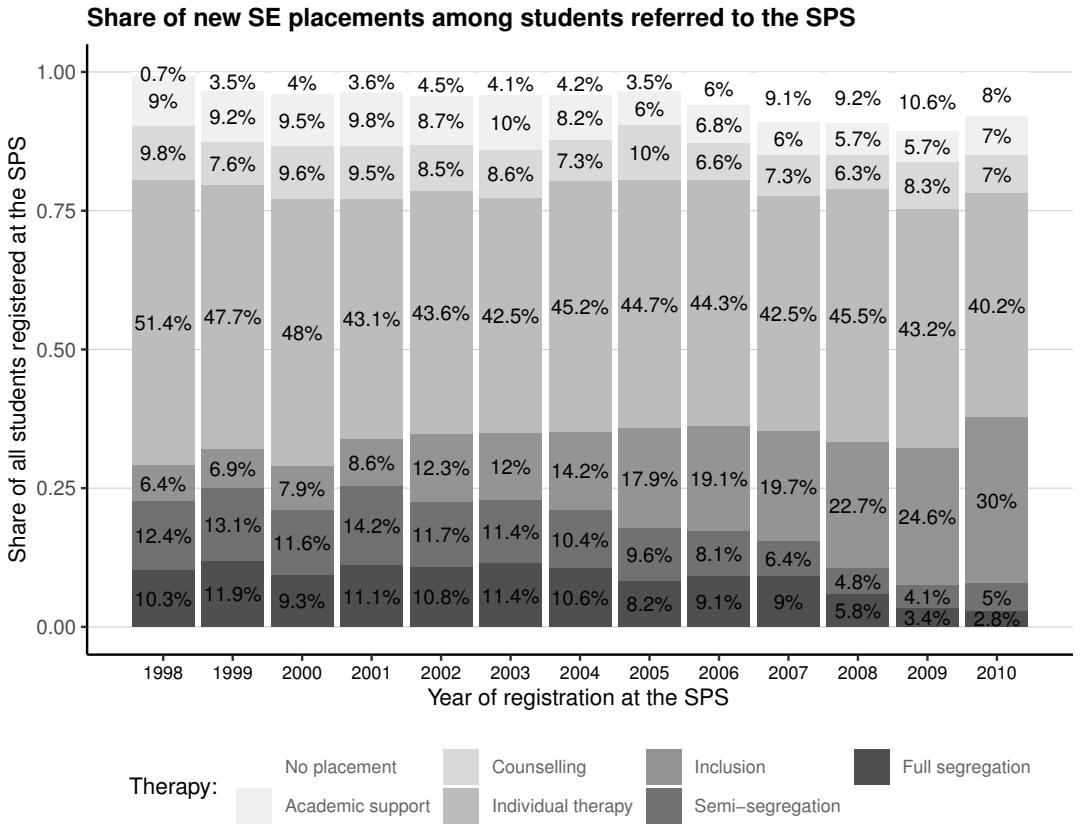


Figure 1: Share of Special Education placements among students referred to the School Psychological Service over the years

Notes: This figure displays the newly assigned special education interventions per year. It gives the share of students who have been referred to the School Psychological Service by special education program over the years. *Source: SPS*.

number of students assigned to small classes steadily decreased (from 12.4% to around 5%). These figures reflect the actual number of SEN students being taught in a semi-segregated settings at the primary level ("stocks"), which dropped from 9.17% in 1999 to 6.4% of all SEN students in 2009, as documented by the official placement register data. Note that each year, a share of students referred to the SPS has not been assigned to any treatment.

The School Psychological Service (SPS) is responsible for the diagnosis and SpEd placement of students with SEN. The SPS is an external and independent administrative entity, and 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. Diagnoses are always made by SPS psychologists, and not by parents, teachers, or school administrators. 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 the requests to the SPS are made when the student is in Kindergarten/Preschool (see Figure C.2). The end of Kindergarten is the moment when teachers decide whether the student

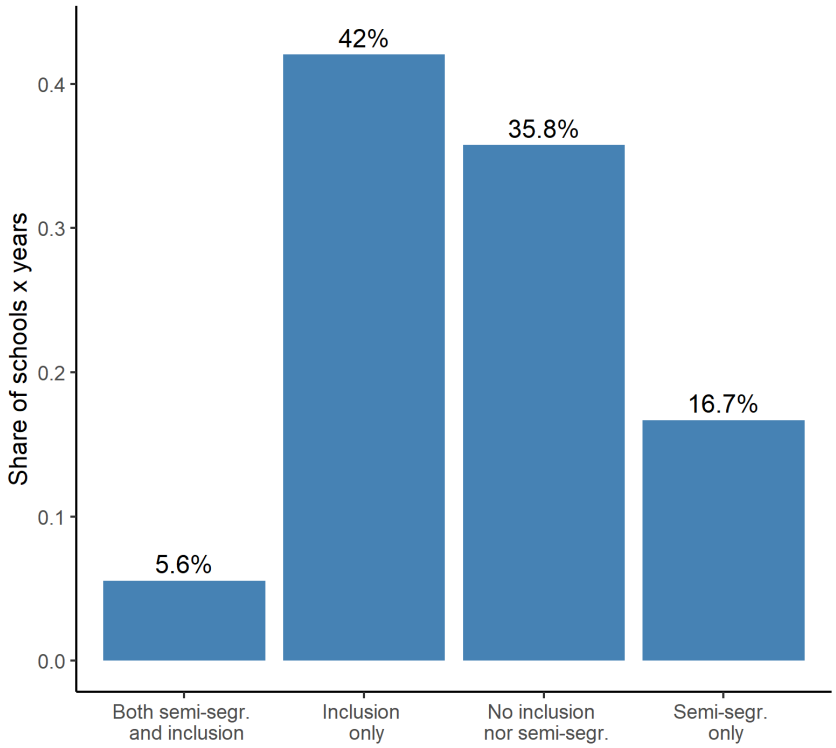


Figure 2: Inclusion and semi segregation across school years

Notes: This figure depicts the fraction of school-by-year units offering semi segregation programs (KK), inclusion programs (ISF), or both/none. One school-by-year unit is one school during one school year, and there are approximately 1326 units (102 schools over the years 1998-2010). Many schools have changed their Special Education strategies over the years. *Source: Pensenpool*.

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.

Each municipality (or groups of municipalities with a shared school) is in charge of setting up their SpEd policies: they choose on a yearly basis which programs to offer among the programs in the basic offer of the Canton. Schools vary substantially in the therapies they offer, as well as in the extent to which they implement inclusive schooling. For instance, Figure 2 shows the fraction of schools-by-years units offering semi-segregation programs (KK), inclusion programs (ISF), or both/none: forty-two percent of all school-years offered only inclusive programs and 17% offered only semi-segregated classes. Finally, students in St. Gallen are assigned to schools on the sole basis of their location of residence. Thus, parents and students must comply with the assignment to the treatment offered by the school.[11]

[11]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).

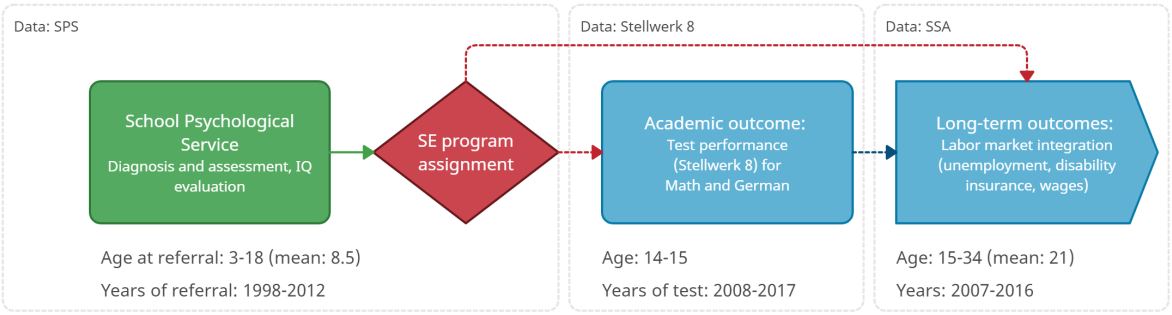


Figure 3: Visualization of the sample structure

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 Stellwerk8 data 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).

2.2 Data: main variables and summary statistics

The main data source on SEN students are the administrative records from the SPS, test scores from the "Stellwerk8" standardized test (SW8) taken in grade 8, and data on labor market integration provided by the Swiss Social Security Administration (SSA). Figure 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 covariates about the student's characteristics, the therapy assigned, the number of visits to the SPS, and the entirety of the psychological records written by the caseworker. All summary statistics are reported in Table 1, and more detailed statistics per treatment status are given in Table C.3 (columns are ordered from the most inclusive program to the least inclusive program).[12]

Characteristics of students are presented in Panel A of Table 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 only 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 correlated with the stringency of the program (more contacts for 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

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

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 education in this paper relies mostly on the text contained in the student-level psychological records written by caseworkers. In many studies, variables that explain treatment assignment are often not observable. Here, the assignment process becomes observable in the text written by the caseworkers. For each visit to the SPS, the caseworker in charge of the student documents the visit, the discussion and following recommendation for SpEd placement. Most comments are quite detailed and offer a comprehensive picture of the problems addressed in the discussion, family background, psychological issues, the diagnoses of the student, and the particularities of the case. These written records provide valuable information as they are written by independent psychologists and not by teachers.

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' mental health, learning/behavioral disabilities, and other background information as much 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 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 summarizes the computational apparatus I use to extract information from text. To avoid making my 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 that learns 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 A.

Program assignment SpEd programs of interest are defined as the programs figuring in the cantonal offer mentioned in Table C.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). Although introductory classes (or "bridge year") between kindergarten

	Mean	Sd	Min	Max	N. obs
A: Individual 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: other	0.043		0	1	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
Regional office: G.	0.099	0.308	0	1	17,822
Regional office: RJ.	0.138	0.356	0	1	17,822
Regional office: R.	0.146	0.356	0	1	17,822
Regional office: Ro.	0.134	0.338	0	1	17,822
Regional office: S.	0.181	0.382	0	1	17,822
Regional office: Wa.	0.136	0.328	0	1	17,822
Regional office: W.	0.163	0.371	0	1	17,822
B: 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 (ISF)	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
C: Outcomes					
SW8 in SW8 cohort	0.763		0	1	13,890
SW8 composit score (SW8 cohort)	0	1	-3.789	4.280	10,602
Used disability insurance (SSA cohort)	0.075		0	1	11,979
Used unemployment insurance (SSA cohort)	0.234		0	1	11,979
Income (std, SSA cohort)	0	1	-1.849	6.855	11,979
D: 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

Table 1: Summary statistics

Notes: 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 2012. *Source: SPS*.

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) (for propensity score only)	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

Table 2: List of used methods for text information retrieval

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 A.

and primary school are SpEd interventions, they are given before primary school and might be followed by further interventions.[13]

Some students (around 5%) were referred by their teachers to the SPS but did not receive any SpEd intervention. These students form an interesting control group, since they are students who raise 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 decided no further intervention was needed. The comparison between students who received a treatment and students who were referred but received no treatment is thus an interesting and meaningful comparison.

Outcomes I measure different outcomes to capture school achievement as well as labor market integration. Outcomes are reported in Panel C of Table 1. For academic performance, I use test scores from the "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 and is the same in all schools. The test 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 SEN students in fully segregated settings are not required to take the test and can choose to opt out, I create an indicator to account for attrition. Around 76%

[13] I consider assignment to introductory classes as a covariate instead of a treatment. Children assigned to introductory classes are reevaluated upon entry in primary school by psychologists and assigned to the programs if needed. Note that many students having had bridge years do not need further interventions.

of SEN students subject to the test actually take it, which is significantly less than the coverage for non-SEN students.

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 most recent income recorded standardized over birth years. This gives the relative position of individual income and accounts for cohort as well as year effects. 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 C.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[14], and students who received more than one treatment. This ensures that multiple influences of different therapies are not confounding the main treatment.

Cohorts registered in the school data and cohorts registered in the SSA data do not perfectly overlap (see red arrows in Figure 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 D of Table 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. Around 46% of the sample belongs to both subsamples. In the results, I conduct attrition analyses and robustness checks using this subsample only.

3 Empirical strategy

Plausible causal estimates of returns to SpEd for each of the five programs mentioned in Table 1 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. I am able to observe the assignment process from an unusually exclusive and detailed perspective and to know almost as much as the caseworker knows about the treatment assignment.

[14]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 to the main measures are, however, kept in the dataset.

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 $\mathcal{X}$ the set of *confounding* variables that are used to account for selection bias, and as $\mathcal{Z}$ the subset of $\mathcal{X}$ that contains the variables used to conduct heterogeneity analysis. The generalized 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, hard 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 with 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 $\text{CATE}_{d,d'}(z) = E[Y_i^d - Y_i^{d'} | Z_i = z]$ where $Z_i \in \mathcal{Z}$. I consider two different cases of CATEs: first, Group Average Treatment Effects (GATEs) give the ATEs for predefined and policy relevant groups of SEN students. For instance, I investigate whether treatment effects are heterogeneous for students who exhibit 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 confounders as heterogeneity variables. This is expressed as $\text{IATE}_{d,d'}(z) = E[Y_i^d - Y_i^{d'} | X_i = x]$.

3.2 Identification

In order to identify returns to SE, the following identifying assumptions must hold. The first key identifying assumption is unconfoundedness, i.e. that there are no features other than X 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. Moreover, unconfoundedness is particularly plausible for the comparison between students placed in semi-segregation and in inclusive

programs. As mentioned above, these two programs are considered as substitutes by the Canton of St. Gallen and each school varies in the extent to which they rely on inclusive programs.

My main identifying assumption relies on the richness of text. To support this assumption, I show that the information retrieved from the text works well to identify students who are similar in observable characteristics. I train different classifiers with the same machine learning methods I use in my main specification to predict students' characteristics that were not extracted from the text. The fraction of missclassified covariates from the text is depicted in Figure C.1 of the Appendix for each text retrieval method presented in Table 2. I find that my text methods are able to capture students' main traits in a reliable way. For instance, my text accurately predicts learning difficulties as a reason for referral for 88% of students, and nonnative status for 90% of students. Gender seems to be the most difficult covariate to predict from the text, as my best text methods capture gender accurately in 70% of cases (which remains a good accuracy score). I also show that my text variables can predict students' IQ score with a very low Mean Absolute Error. In addition to this predictive exercise, I show that text brings additional information which is richer than the information contained only in nontext covariates. Appendix A, and more precisely Figure A.1, Figure A.2, Figure A.3, and Figure A.4, provide descriptive evidence that the text allows me to retrieve valuable information which is not contained in nontext covariates. These figures also show how text information is related to treatment assignment.

The second identifying assumption states that confounders are exogenous, i.e. confounding variables in $\mathcal{X}$ (and in $\mathcal{Z}$) are measured before treatment assignment. Whereas this assumption is easily fulfilled for non-text covariates, it requires more scrutiny for text information: 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.

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. In my case, 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. To deal with potential problems of overlap, I present effects for the overlap population in the Appendix B.

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 peer effects in the classroom generates spillovers. 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.[15] 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. On the one hand, schools in St. Gallen do not engage in strategic therapy assignment against additional budget (e.g., Cullen, 2003). On the

[15]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).

other hand, schools budget their SpEd offers according to guidelines provided and monitored by the cantonal central administration.[16]

Under these assumptions, the estimands of interest are identified using the "augmented" weighted estimator (AIPW) score Γ^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:

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

The "tilting function" $h(x)$ defines the target population as a function of the propensity score $p_d(x)$, and $\omega_d(x) = h(x)/p_d(x)$.[17] 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).[18]

All estimands of interest mentioned above are identified as follows:

$$\begin{aligned}
\text{APO}_d &= E[\Gamma^h(d,X_i)], && h(x) = 1 && (2)\\
\text{ATE}_{d,d'} &= E[\Gamma^h(d,X_i) - \Gamma^h(d',X_i)], && h(x) = 1 && (3)\\
\text{ATET}_{d,d'} &= E[\Gamma^h(d,X_i) - \Gamma^h(d',X_i)|D_i = d], && h(x) = p_d(x) && (4)\\
\text{GATE}_{d,d',z} &= E[\Gamma^h(d,X_i) - \Gamma^h(d',X_i)|Z_i = z], && h(x) = 1 && (5)
\end{aligned}$$

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.

3.3 Estimation with Double Machine Learning

The estimation procedure is represented in the stylized workflow of Figure 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 (and are not fold dependent). Once text representations are predicted on the main sample, the sample is randomly split in K folds of similar size (K-fold cross-fitting, see Chernozhukov et al., 2018). On each $K-1$ folds, I train models to predict the "nuisance parameters" $\hat{\mu}(d,x)$ (the conditional expectation of the outcome) and $\hat{p}_d(x)$ (propensity score) using the covariates $\mathcal{X}$ presented in Panel B of Table 1 as well as text covariates. The nuisance

[16] For each school-year, the cantonal administration determines a target amount of therapy-hours ("Pensenpool") calculated on the basis of the municipality's number of students, its socio-economic "score", and the type of school. Within this given amount of therapy-hours, schools allocate SpEd programs according to their preferred strategy. Schools are obliged to satisfy demand, and for this reason 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.

[17] Note that ω 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).

[18] 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.

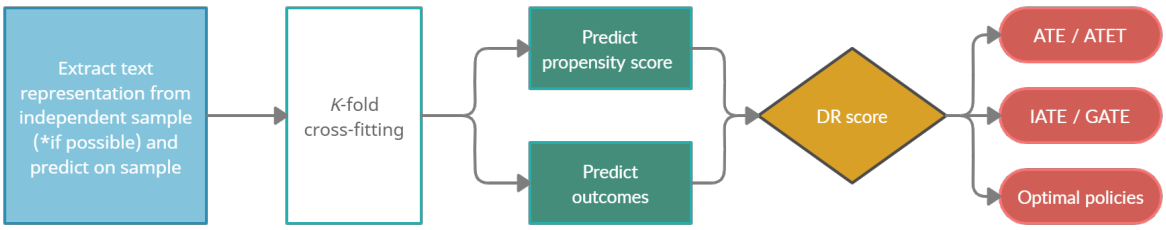


Figure 4: Workflow of Double Machine Learning and text analysis

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 nuisance parameters. The doubly-robust score is computed and used to estimate the estimands of interest (APO, ATE, ATET, IATE and GATE, optimal policies).

parameters are then predicted on the Kth fold. 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.[19] In this 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 used to build, on each left-out fold, the doubly-robust (DR) score as:

$$\hat{\Gamma}_{i,d}^{h=1} = \hat{\mu}(d,X_i) + \frac{\mathbb{1}(D_i = d)(Y_i - \hat{\mu}(d,X_i))}{\hat{p}_d(X_i)}. \tag{6}$$

Since no observation is used to estimate its own nuisance parameter, cross-fitting reduces the risk of over-fitting. 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 (6), the ATE is constructed as the mean of the difference between the APO scores for the treatments of interest, i.e. $\widehat{\text{ATE}}_{i,d,d'} = \hat{\Gamma}_{i,d}^{h=1} - \hat{\Gamma}_{i,d'}^{h=1}$. For the ATET, the doubly-robust score for $\widehat{\text{ATET}}_{i,d,d'}$ is $\left[\frac{\mathbb{1}(D_i=d)(Y_i-\hat{\mu}_{d'}(X_i))}{\hat{p}_d} - \frac{\hat{p}_d(X_i)}{\hat{p}_{d'}(X_i)}\frac{\mathbb{1}(D_i=d')(Y_i-\hat{\mu}_{d'}(X_i))}{\hat{p}_d}\right]$ where $\hat{p}_d = P[D=d] = N_d/N$ (Farrell, 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.[20] The GATEs are estimated by taking the conditional mean of the $\widehat{\text{ATE}}_{i,d,d'}$ over groups determined by

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

[20] 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).

pretreatment variables Z_i, i.e. by estimating an OLS with the score $\widehat{\text{ATE}}_{i,d,d'}$ as outcome variable and standard heteroscedasticity robust standard errors (following Semenova and Chernozhukov, 2020). 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).

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 high propensity score accuracy (Weld et al., 2020) and overcomes difficulties of matching on both covariates and text.[21] 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.[22]

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, and look at which students might still benefit from semi-segregated settings. Finally, I perform a series of further analyses and robustness checks.

4.1 Returns to Special Education programs in inclusive school settings

I first present 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, and individual therapies) or are provided in the main classroom (inclusion) in Figure 5. 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). Effects account for all observed confounding in covariates (such as gender

[21] 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).

[22] 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.

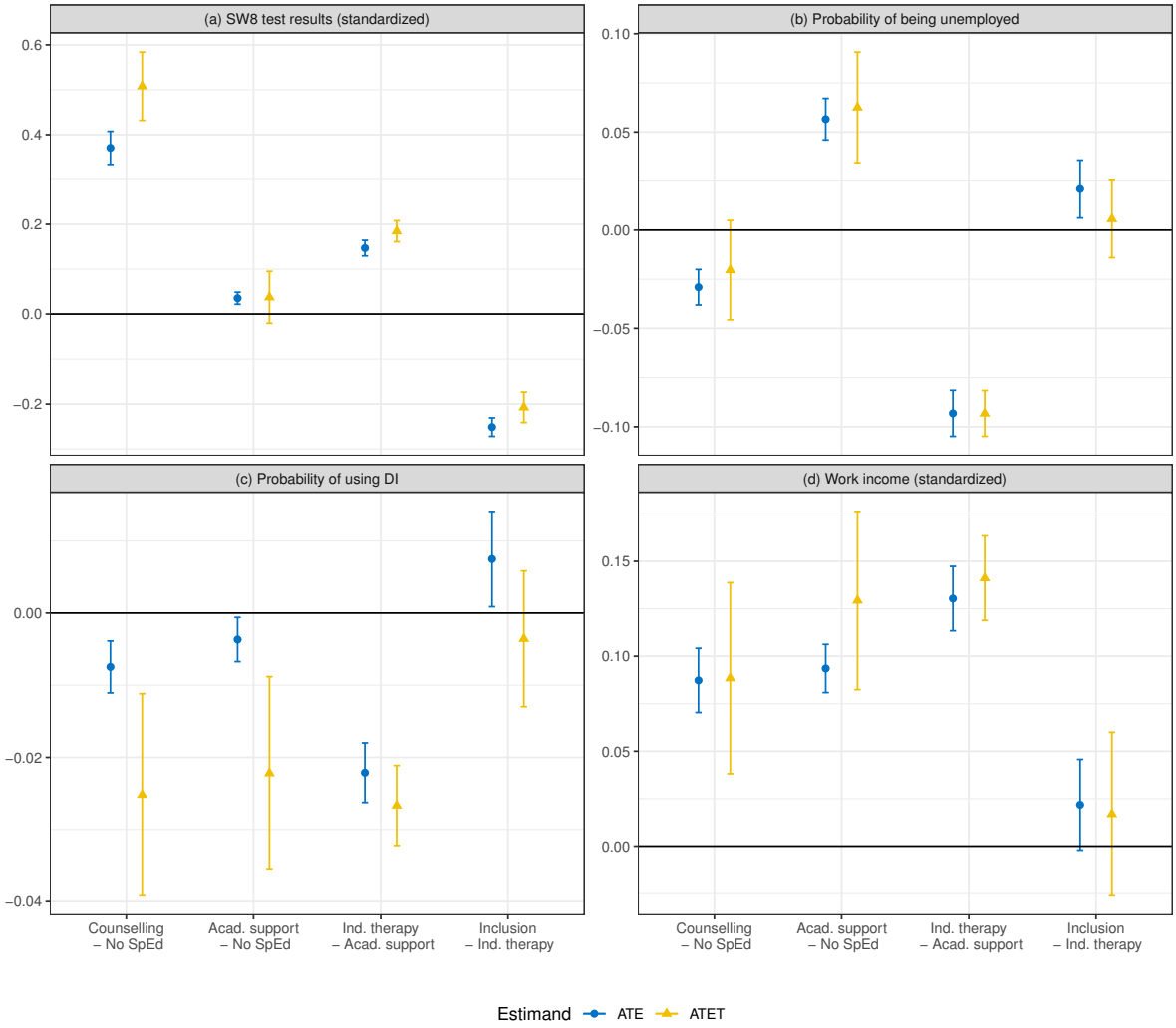


Figure 5: Pairwise returns to Special Education programs according to their level of inclusion.

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. *Source: SPS*.

and IQ) as well as all information contained in the psychologists' written notes. Point estimates and 95% confidence intervals are shown graphically, and both the effect for the whole population (ATE) and the effects for the population of the treated (ATET) are represented.[23] 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 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.

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

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. There are a couple of possible explanations for this difference: first, the way inclusion is implemented varies across schools, and the extent to which students assigned to inclusion are supported by SpEd teachers can differ from individual support to group support. Second, 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.

Turning to labor market outcomes, we see long-term effects that are consistent with the effects on academic performance. As regards the probability of being unemployed and of benefiting from unemployment insurance, results show that counseling (-3 percentage points) and individual therapies (-10 percentage points) have a positive effect (the baseline probability of being unemployed is 23.4%). 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 have almost no effect of inclusion in comparison to individual therapies. Most pairwise effects indicate lower probabilities of benefiting from disability insurance (baseline probability is 0.07), with individual therapies showing the strongest effect. The difference between inclusion and individual therapies in terms of disability insurance recipiency is minute (0.8 percentage points for the ATE, 0 for the ATET). Finally, I find no effects in wage returns, as most pairwise effects exceed 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.

Figure C.5 in the Appendix shows that all 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 do not decrease the probability of showing up at high-stake tests close to the age of graduation, which contradicts the findings of Schwartz, Hopkins, and Stiefel (2021), McGee (2011) or Kirjavainen, Pulkkinen, and Jahnukainen (2016). Moreover, Figure C.5 shows that more serious interventions in inclusive settings are as effective as more benign interventions in ensuring that students with SEN take the test.

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 pop-

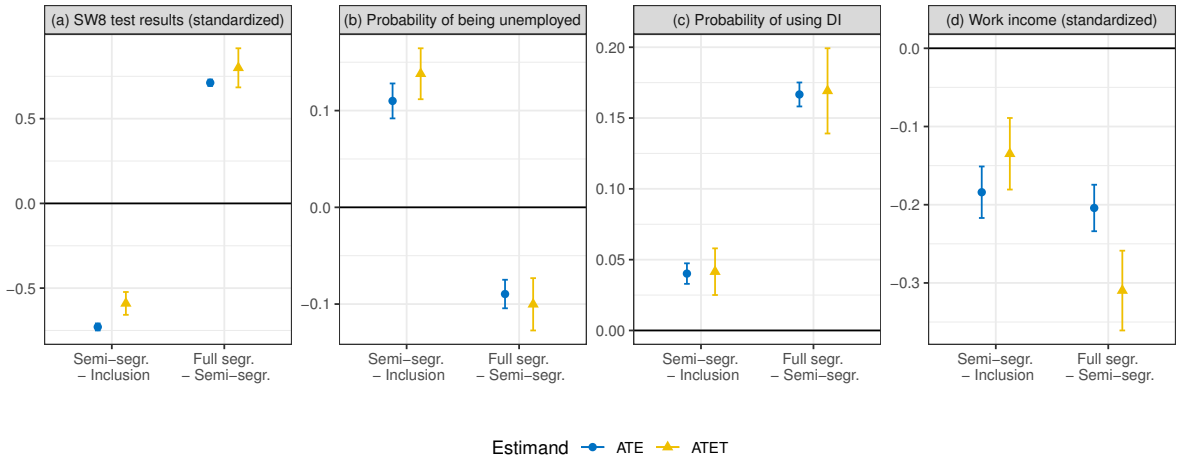


Figure 6: Pairwise returns to segregation.

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. *Source: SPS*.

ulation. 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.

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 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. 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 students segregated as a plausible reason for lower employment. Moreover, these results might be explained by the fact that semi-segregation is attached to a strong 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. This difference is mostly explained by unemployment rather than labor earnings.[24]

Returns to full segregation in comparison to semi-segregation are grim: the relatively positive returns to full segregation in terms of academic performance are mostly due to selection into test participation (see Figure C.5). 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 SE. 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). SEN students are less likely to be sent to full segregation for learning disabilities.

4.3 Returns to Special Education interventions against no intervention

Most of the literature evaluating SpEd programs focuses on returns to SpEd in general without making a distinction between 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 comparison between receiving an intervention and not receiving one becomes more and more difficult to make as interventions become more intensive (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 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 (for inclusion, I measure -0.0878 for the ATE and -0.121 for the ATET). Returns to inclusive programs in term of unemployment are null to positive for all programs

[24] I estimate a specification in which students whose income comes from disability insurance or unemployment benefits are removed from the sample of wages, and I find no pairwise effects of SpEd programs on wages. Results are available on request.

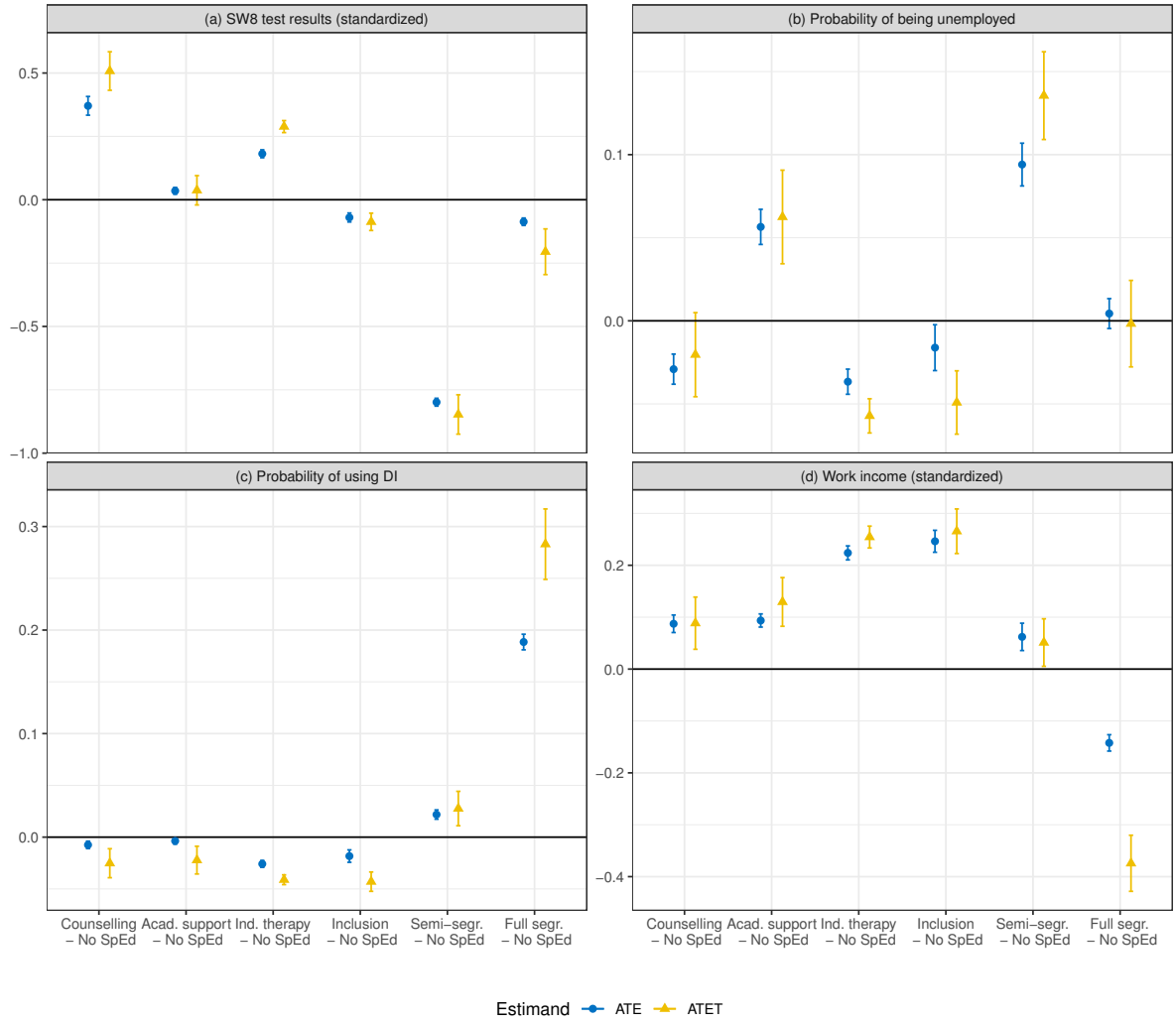


Figure 7: Pairwise returns to Special Education vs. No Special Education.

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*.

except for academic support. Students in inclusive setting catch up on the labor market: they are less likely to be unemployed (Figure 7b.), are less likely to end up on disability insurance (Figure 7c.), and earn as much or even more than students having received no SpEd (Figure 7d.).

Effects of segregation are generally negative. Segregated interventions have negative academic returns (the negative effects of full segregation are mostly 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

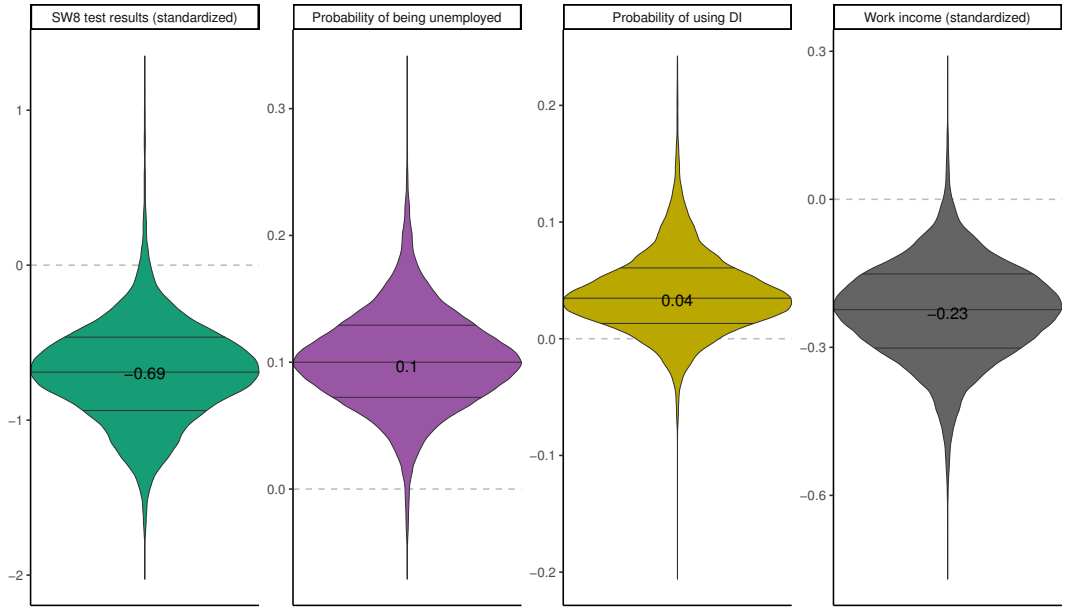


Figure 8: Distributions of IATEs for segregation vs. inclusion.

Notes: This graph represents the distribution of IATEs for the four outcomes of interest. The IATEs were predicted out-of-sample using the DR-learner presented in Equation (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. The smoother is an Epanechnikov kernel.

of benefiting from disability insurance, but not of becoming unemployed. Interestingly, the channels explaining the lack of labor market participation are different for semi-segregated and fully segregated students: whereas students in segregated classrooms are more likely to be unemployed, students in special schools are more likely to become disability insurance recipients.

4.4 Inclusion and semi-segregation: who benefits from segregation?

There is an ongoing debate on whether inclusion is the best treatment for all students with SEN. Despite the policy push towards inclusion, a nuanced analysis on which students would 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 a specific group of students.

4.4.1 Individual effects of semi-segregation

I first investigate Individual Average Treatment Effects (IATEs) for students placed in inclusion vs. semi-segregation. 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).[25] IATEs give the treatment effects at the most granular level and allow the identification of students with SEN

[25] 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 to predict $\widehat{\text{ATE}}_{i,d,d'}$ from 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 over-fitting. 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.

	Test scores			Probability of unemployment			Wage		
	Quint. I.	Quint. V.	SMD	Quint. V.	Quint. I.	SMD	Quint I.	Quint. V.	SMD
Main covariates									
Female	0.45	0.38	0.139	0.41	0.39	0.045	0.48	0.35	**0.282**
Nonnative	0.07	0.34	**0.729**	0.06	0.22	**0.477**	0.04	0.21	**0.552**
IQ	93.41	97.20	**0.341**	94.66	94.82	0.013	95.82	93.66	0.183
Age referral	8.00	8.39	0.188	10.06	8.61	**0.626**	10.06	8.40	**0.752**
Referral: social/emotional problems	0.13	0.37	**0.564**	0.13	0.28	**0.387**	0.13	0.27	**0.343**
Referral: performance/learning problems	0.87	0.90	0.089	0.91	0.79	**0.350**	0.85	0.84	0.029
Referral: conflict with teacher	0.02	0.05	**0.213**	0.01	0.04	0.141	0.01	0.03	0.107
Need psychological treatment	0.13	0.26	**0.310**	0.14	0.24	**0.275**	0.11	0.25	**0.352**
Nonnative × Female	0.04	0.13	**0.339**	0.03	0.08	**0.220**	0.02	0.08	**0.284**
Nonnative×Social/emotional	0.01	0.11	**0.433**	0.01	0.06	**0.309**	0.00	0.06	**0.333**

Table 3: Classification analysis of IATEs for Inclusion vs. Semi-segregation

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. For two SE programs w and w', SMDs are computed as $\frac{\bar{x}_w - \bar{x}_{w'}}{\frac{\sqrt{s_w^2 + s_{w'}^2}}{2}}$, 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. For the probability of unemployment, the first and fifth quintiles are reversed (as students in the fifth quintiles suffer the most from semi-segregation). SMDs higher than 0.2 are depicted in bold.

who benefit the most from each treatment assignment. Figure 8 reports the smoothed distribution of IATEs per outcome with the light bars depicting the first and fifth quintiles of the IATE distribution.[26] The distributions of IATEs are quite spread out, indicating large variations 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, Fernández-Val, and Luo (2018). I group the predicted IATEs into quintiles and compare the standardized means difference (SMD) of covariates for students in the highest and the lowest quintiles. 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 3 shows which students are 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 SEN students 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 better benefit from segregation in terms of labor market integration. Finally, gender

[26]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). Results are very similar, and are available upon request.

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.

4.4.2 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, 2013; Lavy and Schlosser, 2011), students with behavioral problems (Fletcher, 2009), or nonnative speakers (Diette and Uwaifo Oyelere, 2014; Cho, 2012). In addition, I look at treatment heterogeneity along IQ scores.

Findings about the "disruptiveness" hypothesis are presented in Figure 9. Each row of the figure gives the results of a regression where the pseudo outcome is regressed on the group dummy. Red dots indicate the treatment effect for the reference category (those students who do not belong to the group), 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 show clearly 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

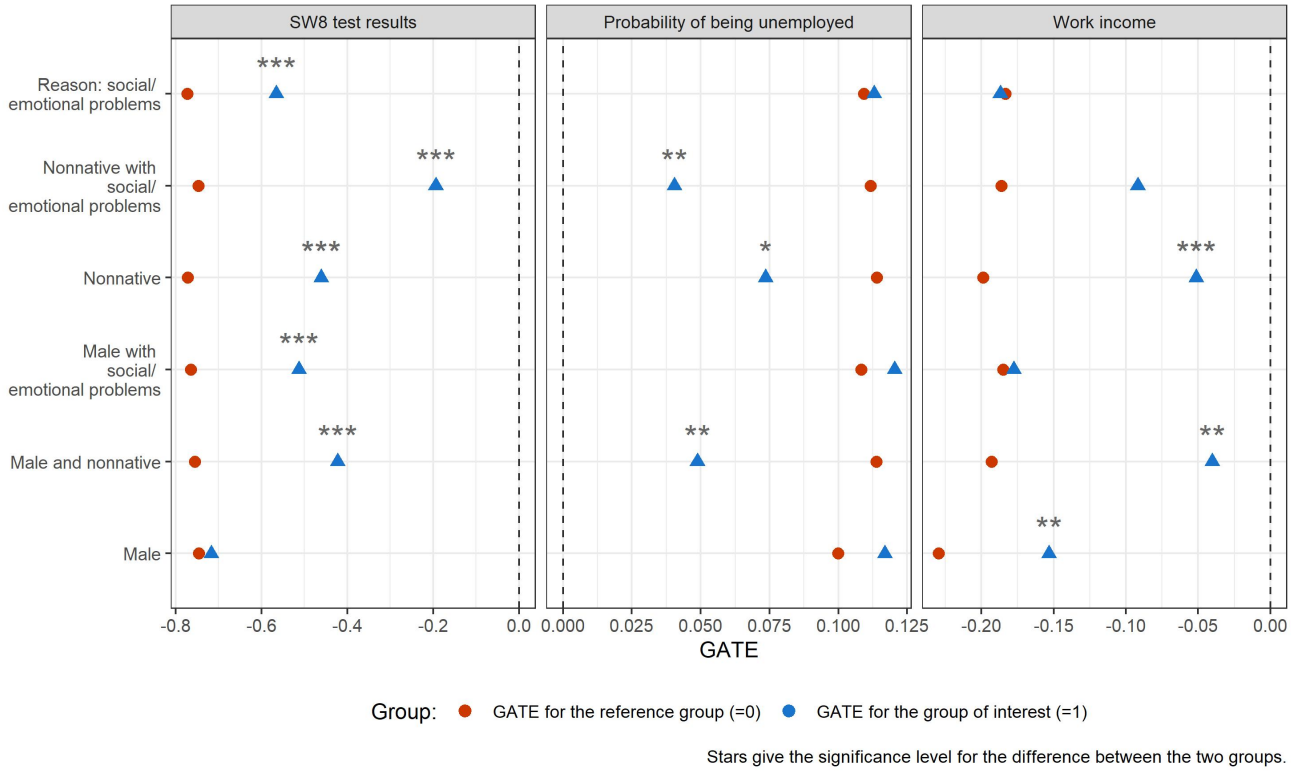


Figure 9: GATEs for semi-segregation vs inclusion

Notes: This graph presents estimated GATEs for the treatment effect of semi-segregation vs. inclusion. Red dots indicate the treatment effect for the reference category (those students who do not belong to the group), 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 ($p < 0.01$).

that would see the smallest difference between inclusion and semi-segregation are nonnative students with emotional or behavioral problems, who would perform only .19 standard deviations less in segregation than in inclusion (and almost 0.6 standard deviations better than other students in semi-segregation). All in all, inclusion remains on average better for students with SEN, even for those with "disruptive" characteristics.

Finally, I explore treatment heterogeneity in IQ scores for inclusion and semi-segregation in Figure 10. To estimate CATEs with respect to IQ, I regress the IQ score on the pseudo-outcome following Zimmert and Lechner (2019) as explained in Section 3.3. In general, there is only minor heterogeneity along IQ in treatment effects of semi-segregation in comparison to inclusion. Students with high IQ face negative returns of segregation that are around half than students with low IQs. For wages, there is no heterogeneity along IQ at all (the line is flat). Students with IQ scores below 85 drive the negative effects of semi-segregation on the probability of using disability insurance. I however remain cautious when interpreting effects for students with IQs lower than 85, as support across treatment groups becomes more difficult to reach (only a few observations drive the effects at levels of IQ below 85). It is also interesting to notice that even students in the higher end of the IQ distribution have lower academic performance and are less likely to be active on the job market.

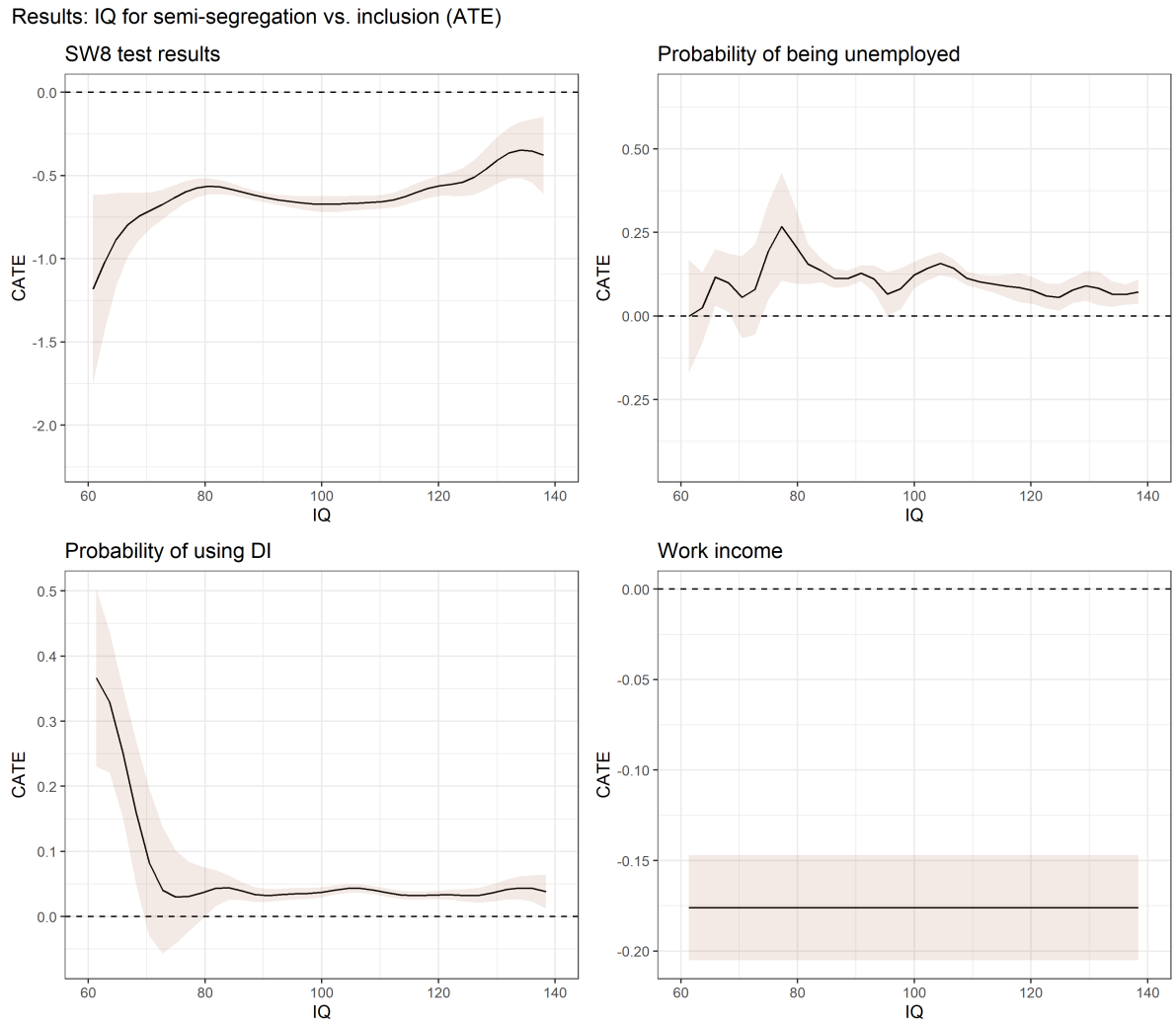


Figure 10: CATEs in IQ for the treatment effect of semi-segregation vs. inclusion.

Notes: This graph depicts the CATE of semi-segregation vs. inclusion along the IQ score for the four outcomes of interest. Inclusion is the reference category. The kernel regression is estimated with a second-order Gaussian kernel function with bandwidth as the 0.9 of the bandwidth chosen with leave-one-out cross-validation. 95% confidence intervals are represented.

In conclusion, my analysis almost undeniably 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 intelligence, gender or the prevalence of learning disabilities: this is especially true for school performance, but also extends to labor market integration.

4.5 Further analyses and robustness checks

I conduct a battery of robustness tests in Appendix B. In Appendix B.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 B.2, I first investigate how sensitive my estimates are to the inclusion of text covariates in order to see how much confounding my text variables 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. Second, I show that the problem of text-induced endogeneity does not harm my estimates. This problem 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. I show that my estimates remain consistent with my main findings when I strip out psychologists' effect from the text.

In Appendix B.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.

In Appendix B.4, I implement an instrumental variable stratgey using instrumental forests (Athey, Tibshirani, and Wager, 2019). 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. 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 variation in school supply of programs 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). Following the Swiss Equality Act for People with Disabilities (2004), municipalities in St. Gallen were strongly encouraged to implement inclusive SpEd programs instead of segregated ones. However, each municipality remained free to offer inclusive measures, semi-segregated ones, or both. LATE results corroborate my main findings.

5 Policy learning: optimal SpEd placement

How would a school psychologist or a policy maker assign SEN students to the program that corresponds the best to their particular characteristics? In this section, 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 observed individual characteristics of interest Z_i and assigns individuals to an optimal treatment d. The optimal treatment is defined as, for each individual, the treatment that maximizes the APO given Z_i, i.e. $E[Y_i^d|Z_i = z]$. For each policy rule $\pi(Z_i)$

among all candidate policy rules Π, the population potential outcome that could be attained is summarized by the policy value function $\hat{Q}(\pi) = E[Y_i^{\pi(Z_i)}] = \frac{1}{N}\sum_{d=1}^{D}\sum_{i=1}^{N}\underline{1}(\pi(Z_i) = d)\hat{\Gamma}_i^d$. This value is computed with the APO under each treatment defined by Equation (6). The goal of the policy maker is to find the optimal allocation of SpEd programs such that the average potential outcome for all treated SEN students is maximized, i.e. by finding the policy rule π^* 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 about how they potentially could improve policies.

For instance, a policy maker could propose three candidate policies for placements: (1) assign all SEN students to inclusion; (2) assign all SEN students to semi-segregation; (3) 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 APO.

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).[27] I experiment with policy trees of depth 2 and 3 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 no unemployment, which is 1-probability of being unemployed at some point). I subsequently link optimal policy allocations to estimated policy costs. I estimate policy costs by taking average costs per student per year of each SpEd placement given by the Canton of St. Gallen and compute the costs of the implemented policy versus the counterfactual costs of optimal policies.[28] I then compare changes in costs and changes in policy value for each optimal policy.

Panel A of Table 4 shows, for each computed policy tree, the percentage of students assigned to inclusion and the percentage of students assigned to semi-segregation. 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 implemented policy is -0.46, and it is 0.47 for labor market integration (which gives the average probability of no unemployment under the implemented policy). I propose four reallocation policies for each outcome. All proposed policies dramatically improve the policy value by reallocating almost all students to inclusion rather than segregation. All proposed policies would reduce overall costs, but not dramatically so (at around

[27] 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).

[28] 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.

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 mio 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 no unemployment. 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

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:***: p <0.01, **: p < 0.05, *: p < 0.1.
Optimal policies computed on 10-fold cross validation.

Table 4: Optimal policies for inclusion and semi-segregation

Notes: **Panel A** of this table shows the treatment assignment from four different policies. The depth indicates the number of tree branches in the policy trees. The baseline variables are individual covariates excluding variables from text, and diagnosis variables are individual covariates + 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** displays validation tests to see whether the proposed policies perform better than either sending everyone to inclusion, sending everyone to semi-segregation, or implementing the (already implemented) observed policy using 10-fold cross-validation. 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.

94% of actual realized cost). In general, proposed assignment schemes are very similar in terms of improved outcomes and reduced costs.[29]

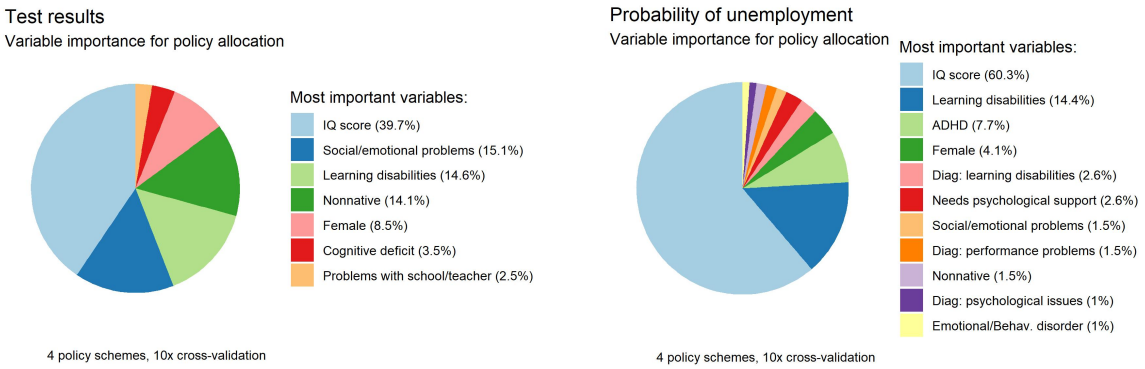
What are the rules that allow better allocation of students in terms of better school performance? Figure C.6 represents the tree policies for improved test scores. Interestingly, a simple policy tree of depth 2 would automatically assign all nonnative students with emotional or social issues to semi-segregation, 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, on the one hand, students with social problems are likely to be disruptive and might benefit from segregation. This policy would outperform the actual assignment by gaining an average of around 0.2 test score standard deviations.[30] 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 4 shows that by sending less students to semi-segregation, the average probability of not being unemployed (which is 1 minus the probability of being unemployed) can be increased by around 20 percent for quasi-similar policy costs. Interestingly, proposed policies that target better labor-market integration send a larger share of students to semi-segregation. Figure C.7 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 "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 is significantly different than zero. Panel B in Table 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 stu-

[29] 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.

[30] 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.



(a) SW8 test scores

(b) Probability of unemployment

Figure 11: Variable importance for optimal Special Education placement

Notes: This graph shows the number of times a variable is split on among all the trees of depth 3 grown in the 10-fold cross-validation. For each variable, the ratio of the number of splits for each variable over the total number of splits is computed. This measure of variable importance shows how much a variable is predictive of optimal outcome.

dents to inclusive settings marginally outperforms the optimal policies for labor market integration; 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 conclude the optimal policy analysis by providing three additional insights. First, I look at which are the most important variables to take into account for a school principal or a policy maker when designing optimal placement policies. This knowledge is particularly important if schools decide to systematically collect information about their students with SEN. To do this, I collect the number of times a variable is split upon for each tree grown in the 10-fold cross-validation. Then I combine the number of splits across all specifications of depth 3 presented earlier (with and without predicted diagnoses) and compute the ratio of the number of splits for each variable over the total number of splits. This measure of variable importance shows how much a variable is predictive of the optimal outcome. Figure 11 shows that IQ is the most important variable (it is split in 39.7% and 60.3% of cases) followed by variables indicating social or behavioral problems as well as learning and performance issues for test score. The presence of learning disabilities and ADHD are important for optimal placement with respect to labor integration. A policy recommendation that follows from this exercise is systematic IQ testing for students with SEN.

Second, I conduct optimal policies for weighted outcomes. So far, I showed that optimal policies might diverge if the policy maker targets different measures of welfare (test scores vs. labor market integration). I construct weighted welfare measures that combine academic performance (with a weight of 70%) and labor market integration (with a weight of 30%). To make both values commensurable, I normalize the APO with a mean of zero and standard deviation of 1. I take the subsample of students who appear in both the SW8 and the SSA datasets. The weighting scheme

is arbitrary, and I chose to weigh academic performance more heavily as this variable is the main outcome of concern for a school official. I present results in Table C.5. Proposed policies would assign 4% to 6% of students to semi-segregation, and the rest to inclusion. They would improve on implemented policies, as well as on fully segregated policies. However, they would be the same or slightly worse than sending all students to inclusion.

Third, I implicitly assumed in my optimal allocation exercise that non-SEN students receive a welfare weight of 0. The reallocation of students with SEN from semi-segregation to the main classroom will induce spillover effects that could negatively impact non-SEN students (as shown by Balestra, Eugster, and Liebert, forthcoming). To measure overall welfare functions, I integrate spillovers of students with SEN in 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 flexible spillover functions of the effect of students with SEN on their peers without SEN for test scores. I replicate the identification strategy and findings by Balestra, Eugster, and Liebert (forthcoming), but I implement more flexible estimation procedures.[31] Figure 12 provides the classroom spillover functions for all classrooms in the Canton of St. Gallen. The function is estimated for students with SEN (in brown) and for students without SEN (in blue). Note that these functions estimate the effect of students with SEN in an inclusive setting, therefore excluding students with SEN in semi-segregated settings. The shape of the spillover function is monotonically decreasing, meaning that including an additional peer with SEN has negative effects on peers, and that the negative effect worsens with more SEN students in the classroom.

The question is: how much harm do we impose on non-SEN students when we reallocate students with SEN to main classrooms? With a back-of-the-envelope utility computation and the assumption that all students have equal utility weight, we see that the increased utility following the inclusion of students with SEN would not offset the utility loss of students in inclusive settings. However, it would harm students already in inclusive settings if students with SEN are not allocated evenly in classrooms. We have 8505 students with SEN in inclusive settings (excluding students receiving semi-segregation, full segregation, and those receiving no SpEd program). These students are spread over 2528 classrooms, meaning that there are 3.36 students with SEN per classroom. A classroom has on average 19.13 students, thus around 17.5% students with SEN. My reallocation policies reallocate 807 students from semi-segregation to inclusion, meaning that on average each classroom receives 0.32 students with SEN (an approximated 1.5 percentage point increase). In classrooms of average SEN composition (17.5% students with SEN), students without SEN have a predicted test score of 0.53, and students with SEN have a predicted test score of 0.27. This is visible in Figure 12.

In the new configuration that includes the 807 students from semi-segregation, students without SEN have a predicted score of 0.51, whereas students with SEN have a predicted score of 0.24. This represents an average utility loss of 0.02 standard test score standard deviations for students without SEN in mainstream classrooms, and of 0.03 standard deviations for students with SEN al-

[31] I estimate spillovers with machine learning algorithms and an ensemble learner. I use a simpler set of covariates, as I do not observe much information about children without SEN (they were never sent to the SPS). I use clustered cross-validation to estimate functions at the classroom level.

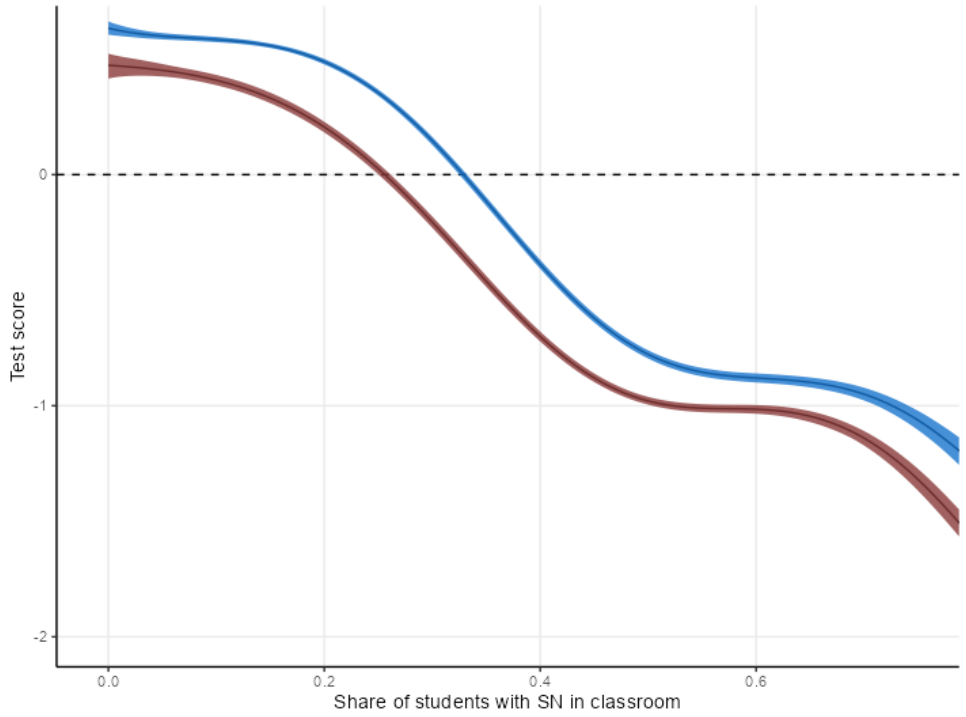


Figure 12: Classroom spillover effects of students with SEN on their peers without SEN.

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.

ready in the classroom. For the newly reallocated students however, the average gains per student are around 0.17 (see reallocation gains in Panel A of Table 4), which by far offset the utility loss of already mainstreamed students. However, the reallocation of one additional student with SEN in a mainstream classroom (now an average 5 percentage point increase) generates negative spillovers of around 0.09 test score standard deviations for students without SEN and of 0.13 test score standard deviations for mainstreamed students with SEN. Although this does not offset the gains for students with SEN, it becomes closer to being equal to the utility gains of reallocated students. This simplistic calculation shows that the utility gain from the reallocation of SEN students from semi-segregation to inclusion would not harm their classmates much if SEN students are allocated evenly across classrooms.

All in all, this simulation 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. We also learn that IQ scores are a measure that can improve SpEd placements. However, predictions on intelligence alone are not enough, and a deeper understanding of the students' disabilities are needed to guarantee effective placement. 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 condition is that students with SEN are allocated evenly. In sum, "inclusion only" policies are a "safe bet" to reach optimal results.

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, I leverage psychologists' written reports to address the problem of confounding. 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. Academic support offers no benefits but does not harm either. When compared to receiving no SpEd, all inclusive treatments have zero to positive returns. I find that inclusion pays off in terms of academic performance, labor market participation and earnings in comparison to semi-segregation.

In general, I find that inclusion is better for (almost) all students with SEN. However, I find that semi-segregation is as good as inclusion for SEN students who exhibit disruptive tendencies. My results 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.

I further explored optimal policy allocations to inclusive and segregated SpEd programs. 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 my proposed optimal policies, a policy maker could significantly increase average school performance and labor market integration at lower overall costs by reallocating students from semi-segregation to inclusion. These clear policy recommendations are interesting in light of the debate on inclusion vs. segregation of students with SEN: by mainstreaming most (if not all) students who would have been assigned to semi-segregation, we can reach higher returns.

This paper invites further research on two fronts. On the one hand, it calls for further research in the field of text as covariate. As mentioned in the Appendix, text as covariate can be highly predictive of treatment assignment, and proper care must be taken to make sure that results are not biased. In this paper, I proposed two "ad-hoc" ways to deal with the problem in my setting, but a more systematic treatment of the problem would be beneficial. Second, a reallocation framework that considers costs, utility gains and spillover effects needs to be developed. In this application,

I addressed these problems sequentially. Solutions to offer a unique and comprehensive algorithm that fleshes out all these decision trade-offs would be handy for researchers and policy makers alike.

References

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.

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 C. 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.

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.

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* .

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.

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

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 N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. " How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star *." *The Quarterly Journal of Economics* 126 (4):1593–1660.

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.

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.

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.

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.

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.

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.

Deuchert, Eva and Martin Huber. 2017. "A cautionary tale about control variables in IV estimation." *Oxford Bulletin of Economics and Statistics* 79 (3):411–425.

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.

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.

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).

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.

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.

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.

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.

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.

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

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

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.

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.

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

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.

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.

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

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

Lazear, Edward 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*, edited by Michael Lechner and Friedhelm Pfeiffer. Heidelberg: 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.

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.

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

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.

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.

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.

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.

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.

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.

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. 2020. "Estimation and Inference about Conditional Average Treatment Effect and Other Structural Functions." *arXiv preprint:1702.06240* .

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.

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

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.

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

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* .

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/.

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.

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* .

Appendix A Using text to adjust for confounding

The purpose of using text extraction methods and Natural Language Processing (NLP) in this paper is for the purpose of 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 high-dimensional and, second, it includes latent features. Because of these two problems, 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 suffering too substantial a loss of 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 problems 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 Text preprocessing

The psychological records for a student contain on average 267 unique words (tokens) and 148 types (unique expressions). However, all implemented methods require the text to be pre-processed — with the exception of word embeddings. For pre-processing, I reduce the text to tokens, strip it from stopwords and lemmatize it.[32] The lemmatization ensures that all tokens are reduced to their stems/roots without inflectional endings. I also all numbers and punctuation signs. Finally, I split the text into onegrams (single tokens) and bigrams (two co-occuring tokens). After pre-processing, the psychological records per student contain on average 241 tokens and 137 types.

A.2 Text representations

Term-frequency matrix (TDM) 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. Thus, two documents are identical (and comparable) if they use the same tokens with a similar frequency. The limitations of TDM is that it does not account for the context in which tokens appear. Moreover, its dimensionality explodes with the number of documents, making comparisons across documents difficult. This requires handling huge sparse matrices, which can be computationally problematic.

[32] For more details on standard practice in text preprocessing, see Manning, Raghavan, and Schütze (2008).

To reduce the dimension of the TDM, I implement two methods: scaling and bounding. For scaling, I weight the TDM 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 selected. The criterion for the choice of scaling scheme and tuning parameters in the *tdm* 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. Figure A.1 and Figure A.2 display the 20 most frequent terms per treatment assignment.

Structural Topic Modeling (STM) and Topical Inverse Regression Matching (TIRM) To mitigate the high-dimensionality and lack of word context in tdm, 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 provides a good comparison of documents, as documents that cover the same topics at the same rates are similar. In this application, I estimate STM setting 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 on the topics that explain topical content the most (see, for instance, Mozer et al., 2020).

TIRM builds on STM by estimating an additional reduction, i.e. a document-level propensity score based on STM and the treatment status as a content covariate. This additional reduction

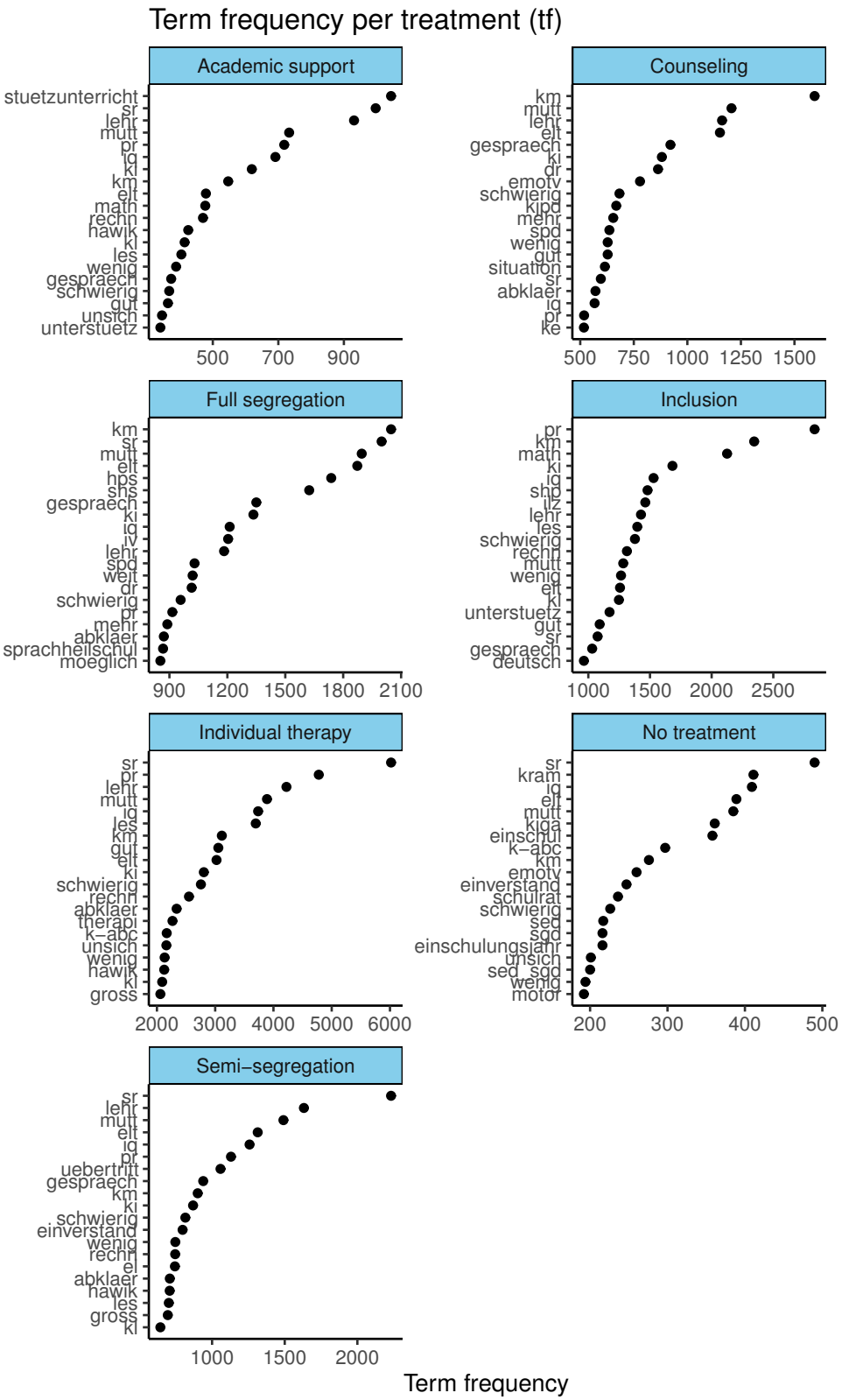


Figure A.1: Most common tokens per treatment assignment

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

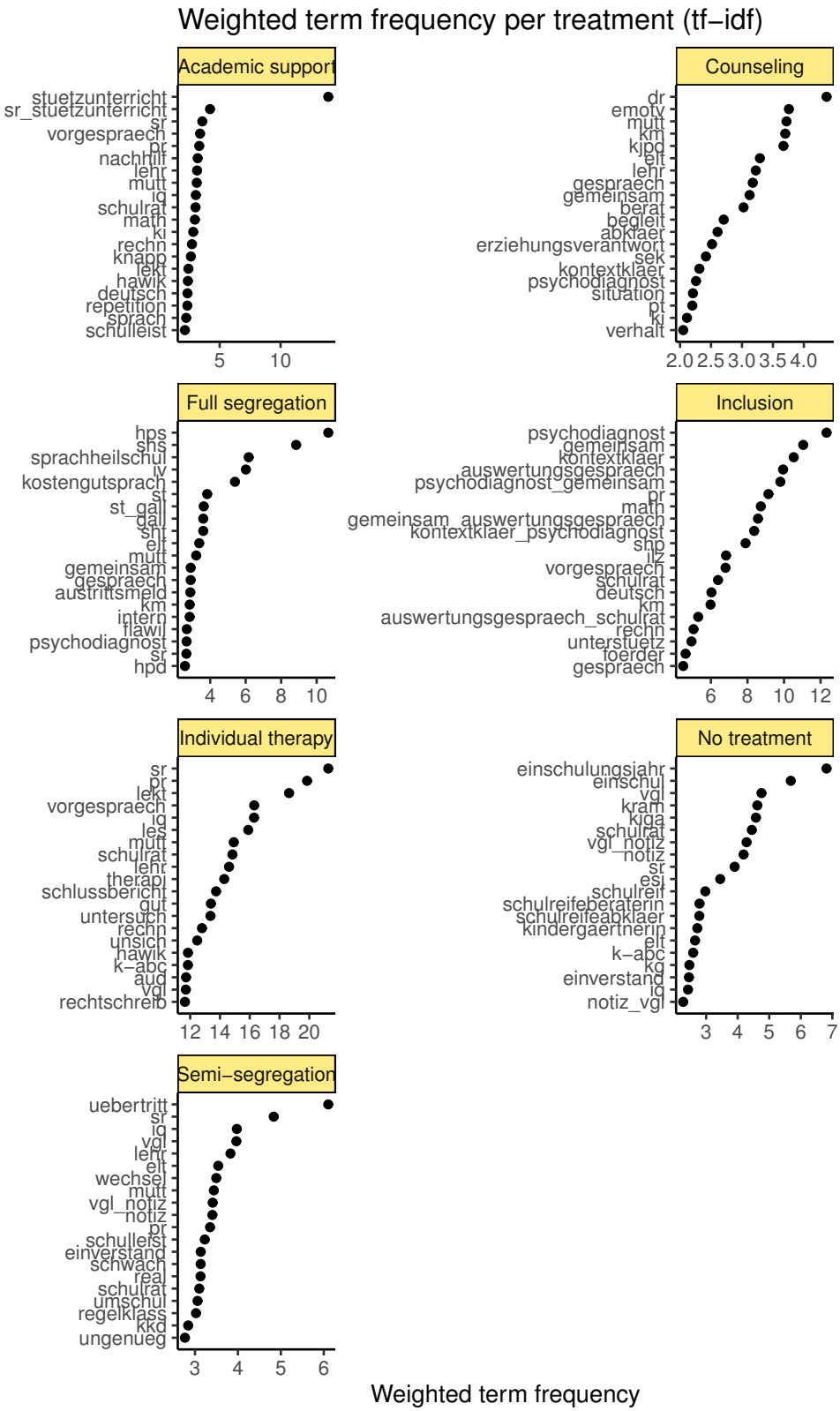


Figure A.2: Most frequent weighted tokens per treatment assignment

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

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, ganterschwil, km, semesterbericht, macht_mueh, agd, shs, notiz, befind, lernverhalt
6	lehr, kjpd, sp, lehrerin, sed_sgd, wld, abklaer, lehr, bess, cpm	rav_pr, luetisburg, 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, legasthenie-schlussbericht, diagnost_termin, wn
8	uebertritt, situation, noetig, therapi, mueh, sv_wld, einverstand, austausch, termin, pl	schlussgespraech_sr, sond, thera, schulpsycholog_abklaer, lektion_woechent, psychodiagnost, sprachheilschul, 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, sprachheilschul, sed, gut, srp	sr_schlussbericht, time-out, z.h_sr, lrsther, textverstaendnis, rt, spezial, antragsschreib_schulrat, thema, slp

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

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 our estimation, similarly to the propensity score estimation, I predict the TIRM sufficient prediction score for each treatment status. I then use the score as additional covariate for the nuisance parameter estimation.

To give an example on how topic modeling can be used to remove confounding, I extract an STM on 10 topics with the above-mentioned covariates 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.[33] Even though STM topics are not always directly interpretable, Table A.1

[33]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 slightly vary with changing samples).

shows interesting patterns. Topic 4, for instance, seems to relate to relationship with school and teachers, whereas topic 6 denotes motor problems or physical impairments. Topic 8 is more about procedural notes and administrative comments. 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 10 are more represented in children who have been segregated in special schools. 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 semantic proximity of words. The algorithm I use is the Word2vec of Mikolov et al. (2013), who proposes 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 only very roughly pre-processed. 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 $N \times W$ vectors matrix. Word embeddings are famous for the word associations they can produce, and I give some illustrative examples in Section A.2. For instance, the words that are the closest to "foreign language", "ADHD" and "dyslexia" are close in meaning to the three expressions.

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 with 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, takes the tokens that are the most frequent (namely, the keywords for each diagnosis). There are many ways to define frequency, and I propose a combination of different measures that are common in computational linguistics.[34] 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 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.[35]

To classify documents into a diagnosis, I first compute the share of dictionary entities per document by dividing the frequency of tokens assigned to a dictionary in a document by the total number of tokens per document. I then weight them such that the sum of each document dictionary frequency totals 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 docu-

[34] This measure is called *keyness*, namely 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).

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

Word	Most similar	Similarity
fremdsprache	subtr	0.66
fremdsprache	groessen	0.64
fremdsprache	rechn	0.64
fremdsprache	einmaleins	0.64
fremdsprache	textaufgaben	0.64
fremdsprache	zahlenstrahl	0.63
fremdsprache	brueche	0.63
fremdsprache	subtraktionen	0.63
fremdsprache	bruchrechnen	0.62
fremdsprache	zr	0.62
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".

ment is discriminatory enough. I hot-encode diagnosis if the proportion of the given diagnosis is 1.5 standard deviations above the mean of diagnosis frequency within 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 rule, and use the same lexicon, texts are quasi-identical.

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 "heavier" 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.[36]

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

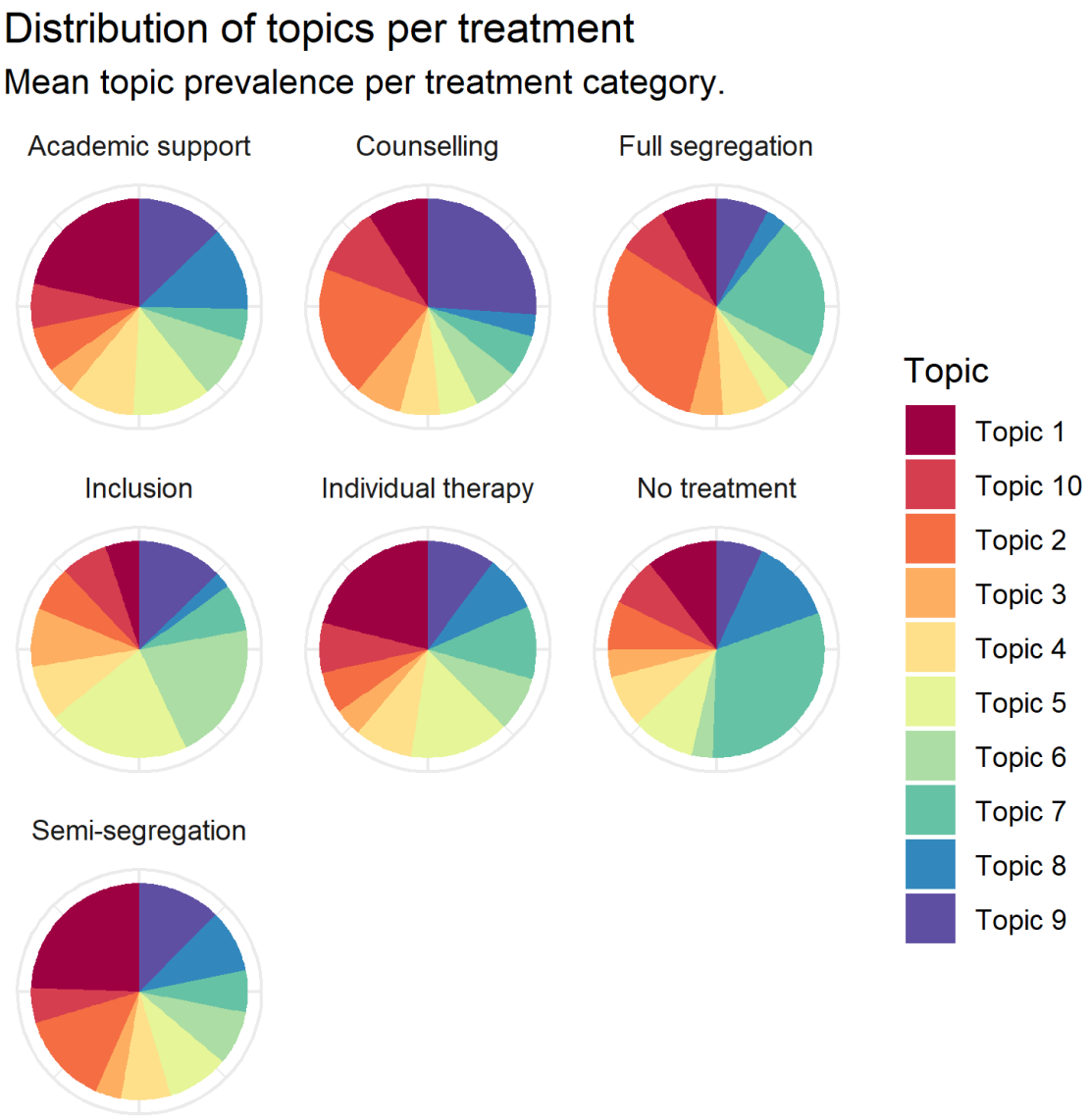


Figure A.3: Prevalence of STM topics per treatment assignment

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*.

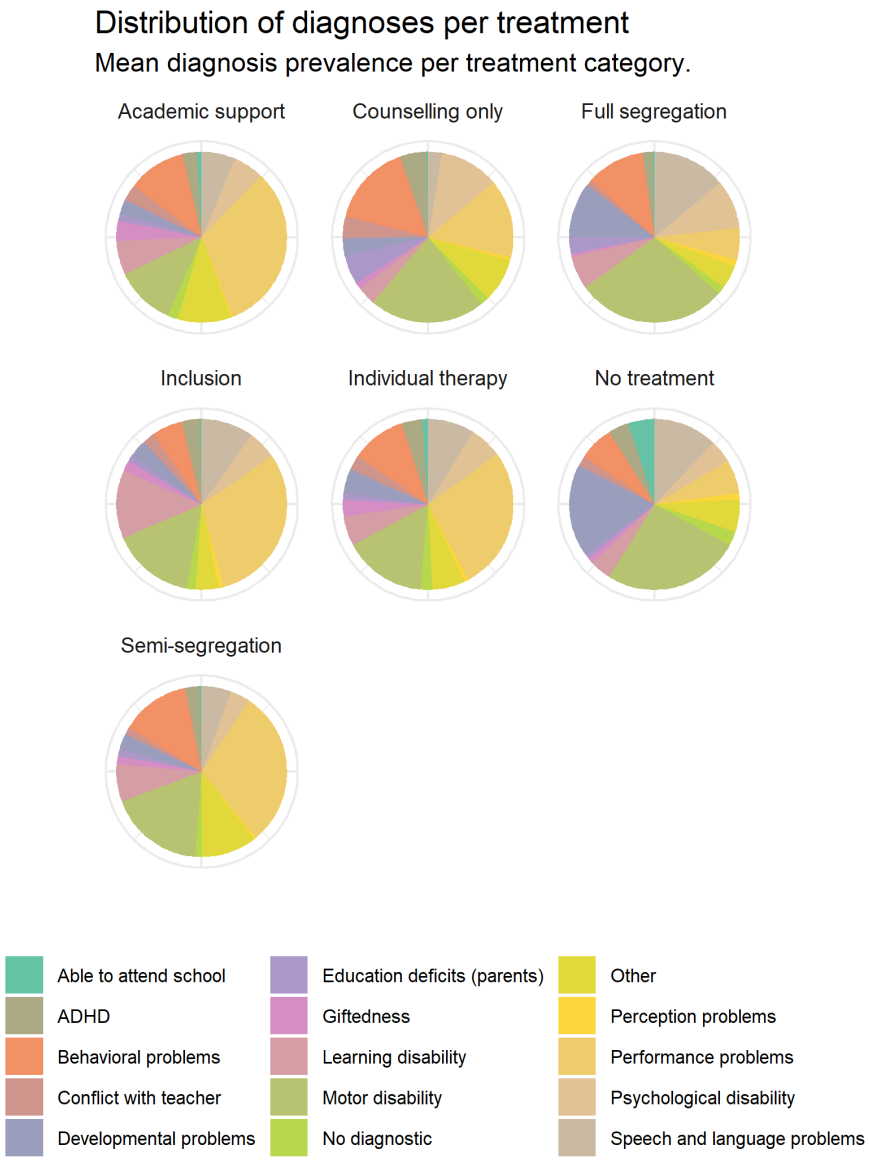


Figure A.4: Average diagnosis per treatment assignment

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*.

Appendix B Robustness and sensitivity checks

B.1 Overlap: ATO, alternative trimming schemes and common support

The problem of overlap is a point of contention and 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), by using different trimming rules and by showing sensitivity to common support.

Average treatment effect on the population of Overlap (ATO) If SN students 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:

$$\text{ATO}_{d,d'} = E_{\text{overlap}}[\Gamma^h(d,X_i) - \Gamma^h(d',X_i)], \quad h(x) = \sum_{k=1}^{D}\left(\frac{1}{p_k(x)}\right)^{-1} \tag{7}$$

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.[37].

The point estimates of the ATO are, in most cases, quite similar to the ATE point estimates, which indicates that there is good overlap in the overall population. This ensures 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 (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 α. 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.

[37] 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`

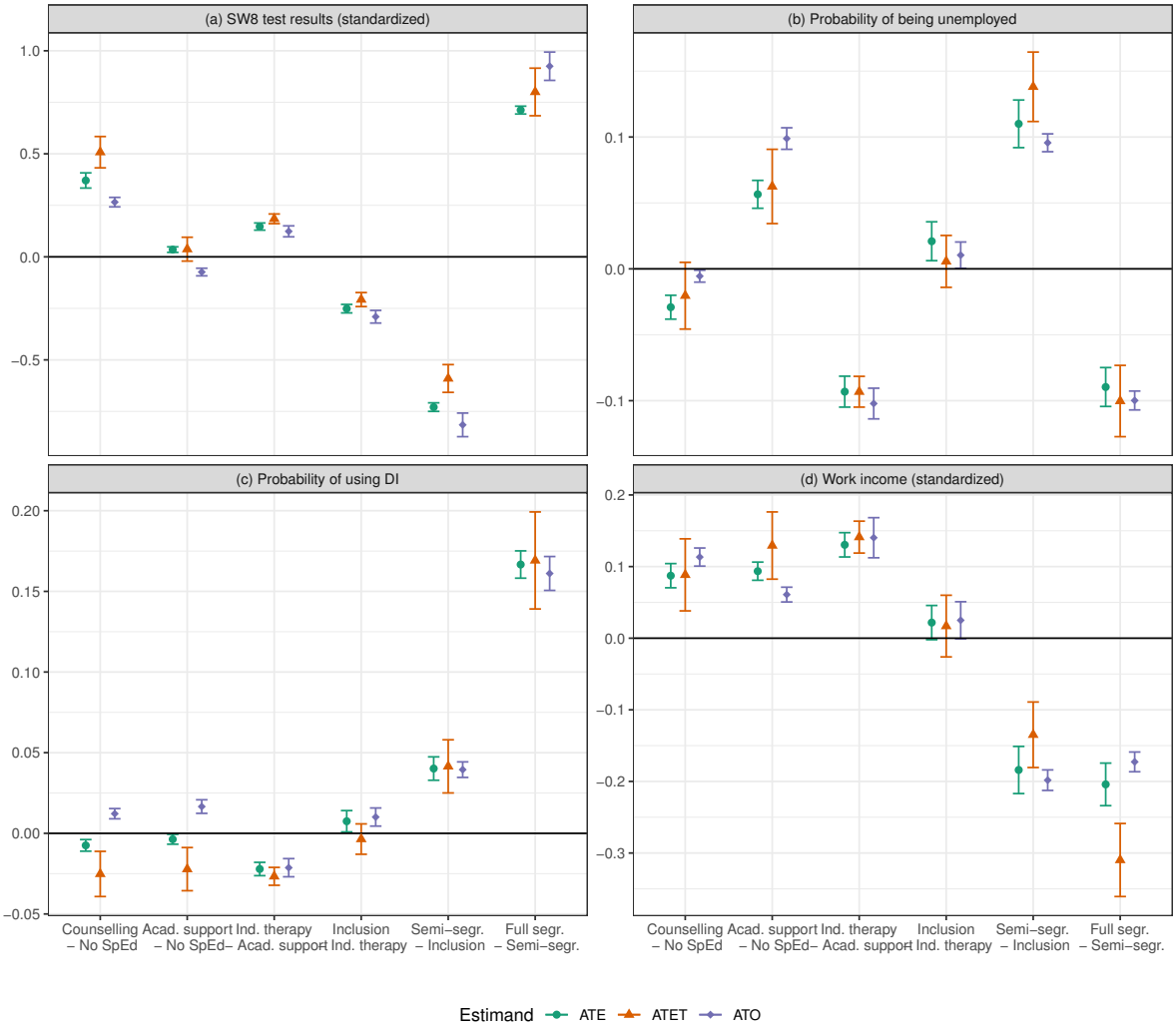


Figure B.1: Pairwise treatment effects with ATO estimates

Notes: This table 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. The treatment effect on the whole population (ATE), the treatment effect on the population of the treated (ATET), and the treatment effect 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 SE" 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*.

To define the trimming schemes and trimming thresholds α, 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 B.2: the majority results discussed in Section 4 persist across different trimming schemes.

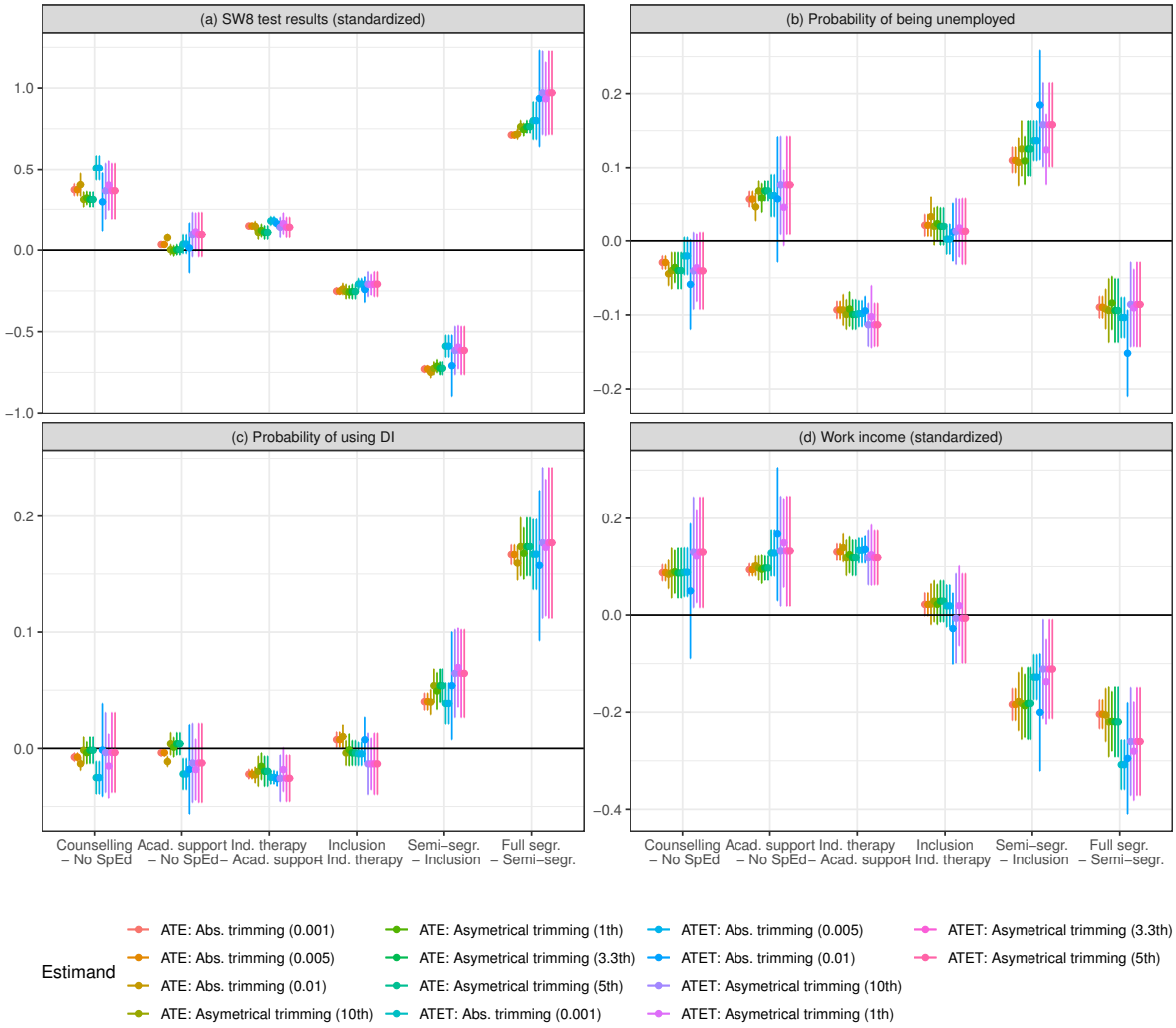


Figure B.2: Pairwise treatment effects with different trimming rules

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 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*.

B.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 how sensitive my estimates are 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 re-compute all my results using only nontext covariates. I then show how much the

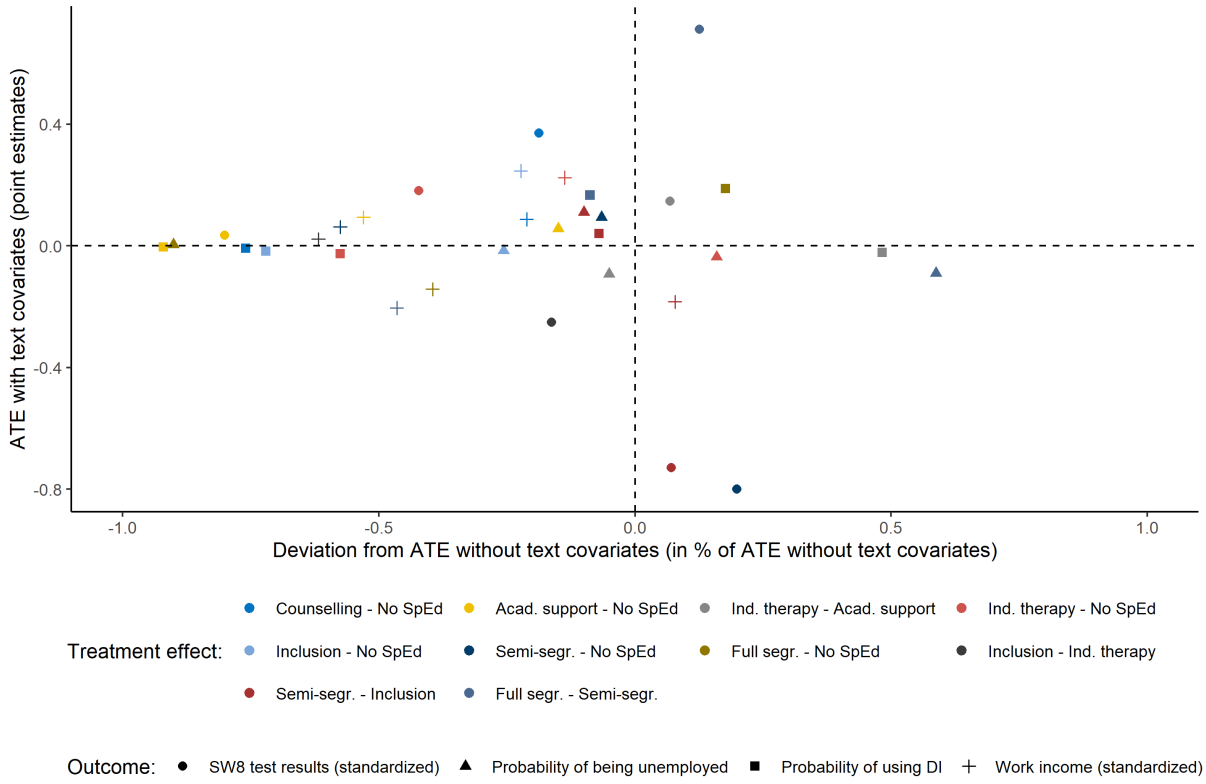


Figure B.3: Effect comparison with specification without text covariates

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.

estimates based on text vary with respect to estimates obtained without controlling for text. This exercise gives me an indication on how much confounding I can remove by using the text.

Figure B.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 without 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 the psychologist always writes using the a set of words-tokens particular to her, the text discovery function g will capture the psychologist's biases together than information on the child.

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 select, for each psychologist, the most frequently and uniquely used word-token (I compute the "keyness" score across psychologists using different measures such as the chi-squared measure, likelihood ratio and point-wise mutual information). I then remove the word-tokens that have a score higher than an arbitrary threshold, and compute my main text measures on this reduced set of features. Results are similar to the main results presented (figures available on request).

B.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 B.4. Results are in line with main results.

B.4 Exogenous placement into inclusive vs. semi-segregated Special Education programs

I turn my attention to remaining unaddressed selection in the estimation of the causal effect of inclusion vs. semi-segregation, and I implement an IV strategy that accounts for variation in schools' preferences for semi-segregated programs. In the Canton of St. Gallen, a basic SpEd program offer is set up by the central government, and each school is free to implement some (or all) of the programs that are part of the basic offer. Following the Swiss Equality Act for People with Disabilities (2004), schools were encouraged to move from semi-segregation to full inclusion of students with SN. However, schools remained free to keep their segregated classrooms open: the decision to move from semi-segregation to inclusion was locally implemented, and raised many concerns for teachers and parents: most arguments express concerns for risks of teachers' overload, incompatibility between non-SN and included SN students, and a lack of efficiency for SN students.[38]

[38]Newspaper articles from the *Tagblatt* newspaper in St. Gallen regularly refer to the problem. See the articles *"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"*.

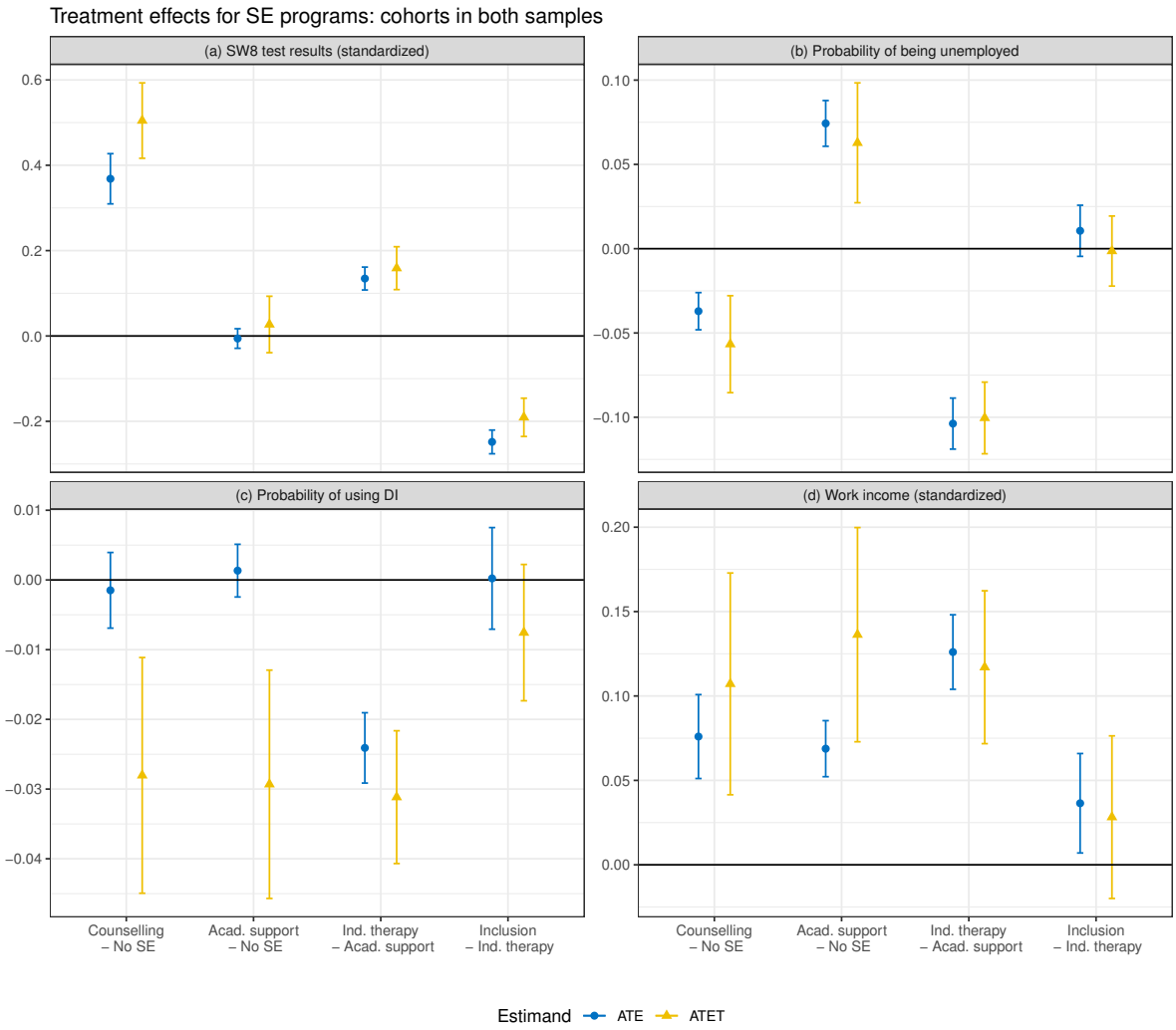


Figure B.4: Pairwise treatment effects for cohorts observed in all outcomes

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*.

I exploit the difference in schools' implementation of segregated vs. inclusive measures as a binary instrument for assignment to inclusion. 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). I collected data on schools' implementation of SpEd programs from the *Pensenpool* held by the ministry of education of the Canton of St. Gallen. These data indicate which school in which academic year has implemented inclusive programs, semi-segregated programs, or both. They also give the number of students assigned to the programs. As I showed

in Figure 2, schools vary in the extent to which they implement inclusive or segregated programs across the years. Some school-years have inclusion only, semi-segregation only, both, or neither. The offering of inclusive programs also varies within schools across years (not shown here).

Thus, local preferences for SpEd segregated programs increase the likelihood that students with SN will be kept in inclusive settings, especially for students with moderate difficulties. The variation in schools' program implementation is exogenous to SN students' characteristics. If the school implements an inclusive program only, students who would have been sent to a segregated classroom are mainstreamed in the normal classroom and given additional support. In this case, compliers are students who would have been assessed and given treatment by the same psychologist but who would have been segregated if they had attended school in another municipality. The IV strategy uses this variation across school-years within school psychological service in the first stage. I report first-stage (linear) estimates in Table B.1. A student being in a school that implements inclusive programs is around 10 percentage points more likely to be enrolled in an inclusive program. These results hold when school fixed-effects, year fixed-effects, caseworker indicators, as well as SPS office fixed effects and individual characteristics are added.

I estimate (Conditional) Local Average Treatment Effects $\tau(x) = Cov(Y_i, T_i|X_i = x)/Cov(D_i, T_i|X_i = x)$ where Y_i is the outcome for individual i, T_i is the binary instrument[39], D_i the binary treatment, and X_i the set of covariates. To integrate the information given by the text, I estimate the score with the instrumental forest of Athey, Tibshirani, and Wager (2019): the algorithm estimates an adaptive weighting function of $\tau(x)$ derived from a forest in order to uncover heterogeneity. As shown in regressions of the first stage, I also include school indicators as controls.[40] I only include students who were sent to either inclusive or semi-segregated programs.

Results are presented in Table B.2. The effect on compliers, i.e. students who would have been assigned to segregation if there were no inclusion amenities in the school, is well in line with ATEs estimated above. Compliers in inclusive settings perform 0.76 standard deviations better in test score, are 15 percentage points less likely to end up in unemployment, and earn significantly more (0.14 standard wage deviations more). Moreover, effects on disability insurance remain null. These estimates reflect my main ATE and ATET estimates, and are robust across different specifications (I experimented with different parameter tuning of the instrumental forest).

The exclusion restriction is violated if the decision for a school to implement an inclusive policy is correlated with the school population or demographic characteristics of the municipality. For instance, school officials could decide to implement semi-segregated measures considering the pool of students in their schools/municipalities. I argue that the instrument is conditionally valid, i.e. that the variation coming from the school's SpEd program is exogenous to the student's characteristics after I control for the school and students' characteristics measured prior to treatment assignment. This is even more plausible given that I observe students' characteristics found in the psychological reports.[41] Moreover, the SW8 test is given in secondary schools, which group students from different municipalities and schools. The test is standardized and is the same for all students across the

[39] I define the instrument as T to avoid confusion with pre-treatment covariates Z_i.
[40] I did not include school indicators in my main results.
[41] See Deuchert and Huber (2017) for a discussion on IV with control variables.

	Probability to be assigned to inclusive program						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
School with inclusion	0.548***	0.530***	0.481***	0.303***	0.289***	0.114***	0.099***
	(0.013)	(0.013)	(0.012)	(0.014)	(0.014)	(0.016)	(0.017)
Female		0.037***	0.028**	0.023**	0.023**	0.022**	0.026***
		(0.012)	(0.012)	(0.010)	(0.010)	(0.009)	(0.009)
Nonnative speaker		−0.184***	−0.187***	−0.119***	−0.083***	−0.061***	−0.040***
		(0.016)	(0.015)	(0.014)	(0.015)	(0.012)	(0.013)
Age at first referral		−0.018***	−0.027***	−0.027***	−0.027***	−0.026***	−0.024***
		(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
IQ							0.005***
							(0.0005)
Share of nonnatives in school					−0.331***		
					(0.053)		
Constant	0.290***	0.477***	0.370***	0.184**	0.194***	0.053	−0.442***
	(0.010)	(0.027)	(0.032)	(0.073)	(0.072)	(0.043)	(0.065)
Year FE	No	No	Yes	Yes	Yes	Yes	Yes
Caseworker	No	No	No	Yes	Yes	No	No
Regional office	No	No	No	No	Yes	Yes	Yes
School FE	No	No	No	No	No	Yes	Yes
First stage F stat	1826.82	1747.82	1566.19	495.22	444.32	51.24	32.92
N	4,189	4,189	4,189	4,061	4,061	4,189	3,359

Notes: Sample with children in ISF and KK only.

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

Table B.1: First stage estimates

First stage estimates of whether a given school has implemented inclusive Special Education programs on the probability for students to be assigned to inclusive programs. Sample includes only SN students assigned to inclusion or to semi-segregation. *Source: SPS, Pensenpool*.

Canton. Furthermore, 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. As regards the argument of potential selection into schools, I refer to existing literature (Balestra, Eugster, and Liebert, 2020, forthcoming; Balestra, Sallin, and Wolter, forthcoming): 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).[42]

Another threat to the exclusion restriction would be that the instrument would influence the so-called "never-takers", i.e. students with SN who would always remain in segregated settings or in inclusive settings. This would be a problem if I included mainstreamed students without SN in my sample, as a marginal increase in the number of included peers generates negative spillover effects (see Balestra, Eugster, and Liebert, forthcoming). However, as I measure the effect of the implementation of inclusive policies on the population of students with SN only, I mitigate this problem: peer

[42]In addition, there is some volatility in "switches" from years with inclusion SpEd programs implementation to years without it within schools, which reduces the risk of parents being able to foresee SpEd program implementation.

IV Estimates: effect of inclusion			
Outcome	LATE	Conf. int. low	Conf. int. high
Only inclusion and semi-segregation			
SW8 test results	-0.757	-0.538	-0.977
Probability of using DI	-0.002	0.073	-0.077
Probability of being unemployed	0.153	0.280	0.028
Work income (month)	-0.145	0.131	-0.421

Table B.2: IV results of inclusion

LATE estimates estimated using generalized random forest with honesty. The forest was grown with 10'000 trees. Text covariates were discovered in a hold-out sample. 95% confidence intervals are given in the table. *Source: SPS, Pensenpool*.

effects in segregated classrooms are likely to be negligible given the low group size, the homogeneity of students, and the high teacher time per student. This argument is likely to also hold for students in inclusive settings, as they also receive additional support from special education teachers.

Appendix C Supplementary Material

Therapy	German	Treatment group
Inclusive special education (ISF)	Integrierte Schülerförderung (ISF), Schulische Heilpädagogik	Inclusion
Special (small) classes	Kleinklasse	Soft segregation
Introductory classes	Einführungsklasse	Intro. classes **(only before primary school)**
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	Supportive measures (not therapy)
Individual learning goals	(reinforcement of ISF)	Supportive measures (not therapy)

Table C.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 C.2: Attrition analysis

	Counselling	Academic support	Individual therapy	Inclusive special ed. (ISF)	Semi-segregation	Full segregation	No therapy (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
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
Income (std.)	-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 C.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	0.24	0.03	0.10	0.21	0.05	0.23	0.12	0.15	0.14	0.25	0.03	0.27
Referred by: Other	0.12	0.08	0.11	0.18	0.06	0.15	0.10	0.03	0.10	0.03	0.22	0.02	0.07	0.06	0.25	0.01	0.12	0.31	0.08	0.20	0.05	0.24
Referred by: Parents	0.17	0.32	0.36	0.40	0.46	0.29	0.41	0.05	0.08	0.15	0.03	0.10	0.04	0.11	0.07	0.06	0.07	0.11	0.02	0.18	0.05	0.13
Referred by: Parents and teacher	0.14	0.08	0.20	0.05	0.01	0.18	0.07	0.11	0.03	0.09	0.26	0.15	0.14	0.21	0.38	0.27	0.06	0.23	0.12	0.17	0.06	0.11
Referred by: Teacher	0.18	0.16	0.01	0.23	0.32	0.23	0.38	0.17	0.07	0.16	0.07	0.22	0.24	0.33	0.24	0.39	0.09	0.01	0.15	0.08	0.06	0.15
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
Chose VET	0.33	0.19	0.15	0.01	0.22	0.37	0.50	0.05	0.19	0.03	0.57	0.71	0.14	0.07	0.52	0.66	0.21	0.37	0.51	0.60	0.74	0.13
Chose no further education	0.32	0.08	0.05	0.14	0.05	0.52	0.59	0.04	0.22	0.04	0.61	0.68	0.18	0.00	0.57	0.64	0.18	0.38	0.45	0.57	0.64	0.07
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.)	0.27	0.03	0.24	0.21	0.06	0.49	0.06	0.21	0.18	0.10	0.54	0.09	0.03	0.31	0.75	0.31	0.27	0.71	0.27	0.42	0.01	0.45

Table C.4: Summary statistics: SMD comparison

Standardized mean differences (SMD) across SE programs. For two SE programs w and w', SMDs are computed as $\frac{\bar{x}_w - \bar{x}_{w'}}{\frac{\sqrt{s_w^2 + s_{w'}^2}}{2}}$, 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*

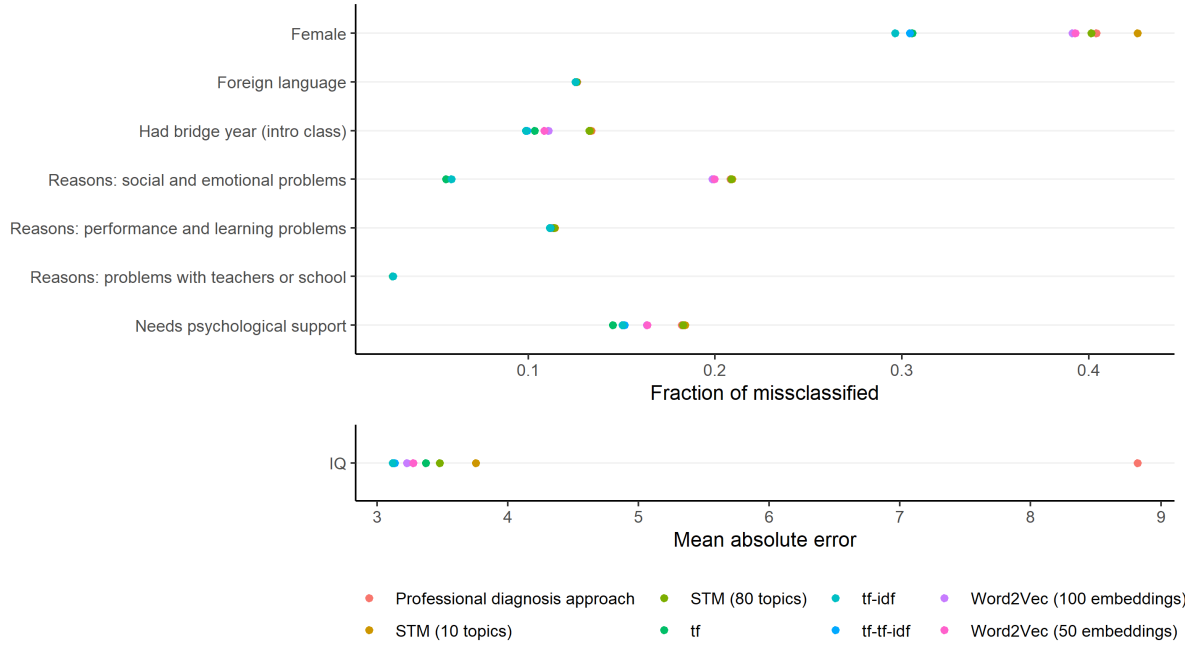


Figure C.1: Prediction of main covariates from text

Notes: This figure represents the fraction of missclassified variables for the main covariates. For each covariate, I train a classifier on my text measures to predict the covariate status. For the IQ, I run a regression and compute the Mean Absolute Error to assess the accuracy of my predictors. Classification and regression algorithms are the methods I use in my main specification (random forest, lasso and elastic net). See Table 2 and Section A for more details on the text retrieval methods used in this figure.

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 mio CHF)	Percent of actual costs
Overall welfare (70% academic performance, 30% labor market integration. N = 1880					
Actual allocation	0.60	0.40	-0.005	40,966	1
Depth 2 and baseline variables	0.96	0.04	0.175	37,942	0.93
Depth 3 and baseline variables	0.94	0.06	0.190	38,140	0.93
Depth 2 and diagnosis variables	0.96	0.04	0.175	37,942	0.93
Depth 3 and diagnosis variables	0.94	0.06	0.190	38,140	0.93

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

	All inclusion	All semi-segregation	Assigned policy
Overall welfare (70% academic performance, 30% labor market integration. N = 1880			
		Optimal policies computed on 10-fold cross validation.	
Depth 2 and baseline variables	-0.011*	0.515***	0.347***
	(0.006)	(0.044)	(0.032)
Depth 3 and baseline variables	-0.013*	0.513***	0.346***
	(0.007)	(0.043)	(0.032)
Depth 2 and diagnosis variables	-0.018**	0.509***	0.341***
	(0.007)	(0.043)	(0.032)
Depth 3 and diagnosis variables	-0.019**	0.507***	0.339***
	(0.009)	(0.043)	(0.031)

Notes:***: p <0.01, **: p < 0.05, *: p < 0.1.

Table C.5: Weighted optimal policies for inclusion and semi-segregation

Panel A of this table shows the treatment assignment from four different policies. The depth indicates the number of tree branches in the policy trees. The baseline variables are individual covariates excluding variables from text, and diagnosis variables are individual covariates + 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** displays validation tests to see whether the proposed policies perform better than either sending everyone to inclusion, sending everyone to semi-segregation, or implementing the (already implemented) observed policy using 10-fold cross-validation. 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.

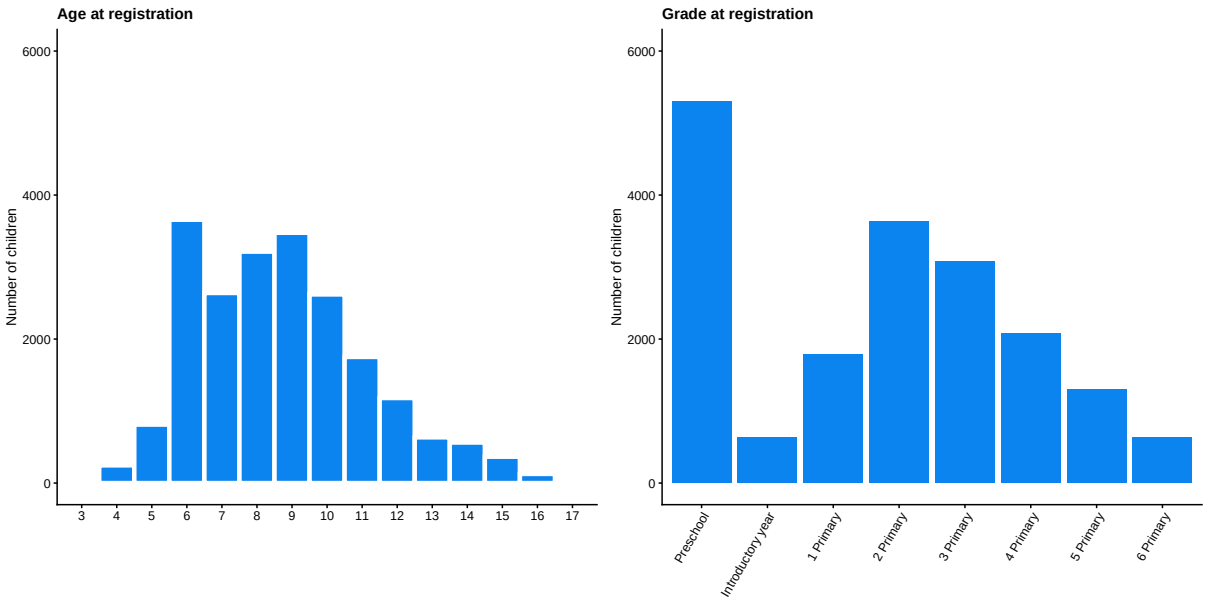


Figure C.2: Distribution of age and grade at registration

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*.

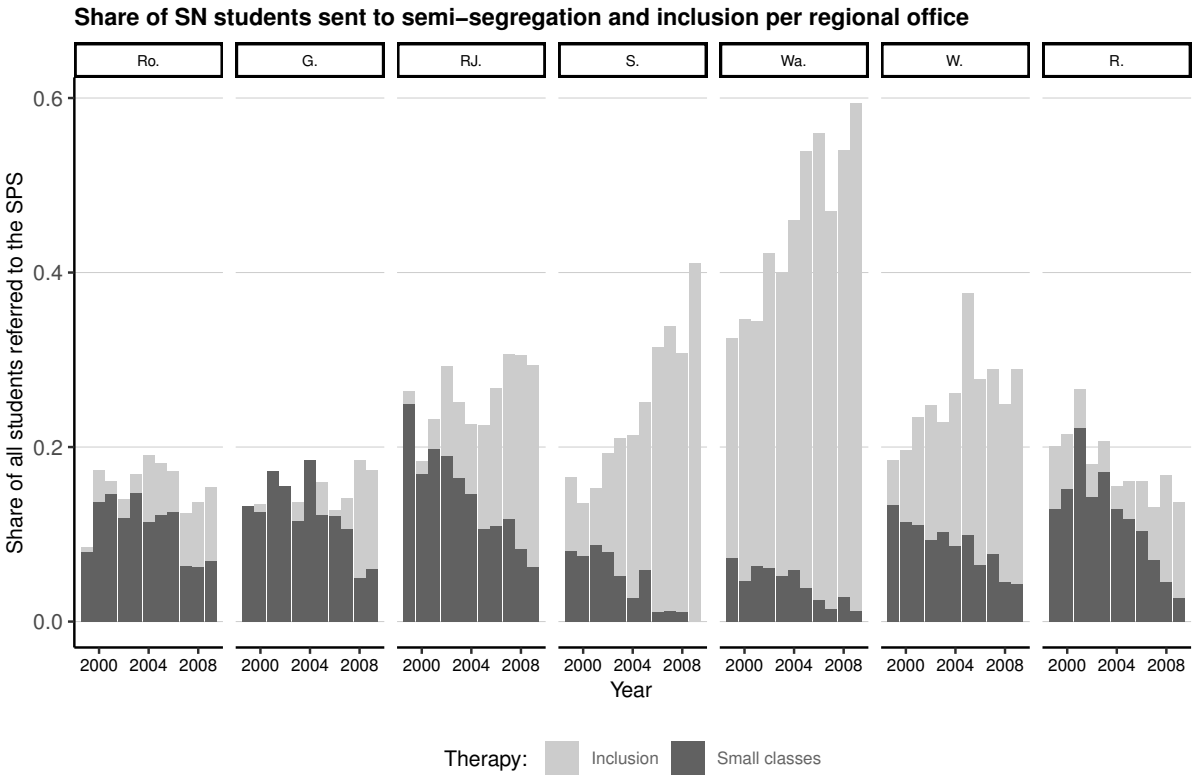


Figure C.3: Share of students sent to semi-segregation and inclusion per regional office

Share of students who have been referred to the SPS of the designated regional office (assigned and non-assigned to a particular therapy). Only inclusive therapy and small classes are represented. The names of Regional offices are abbreviated for confidentiality purposes. *Source: SPS*.

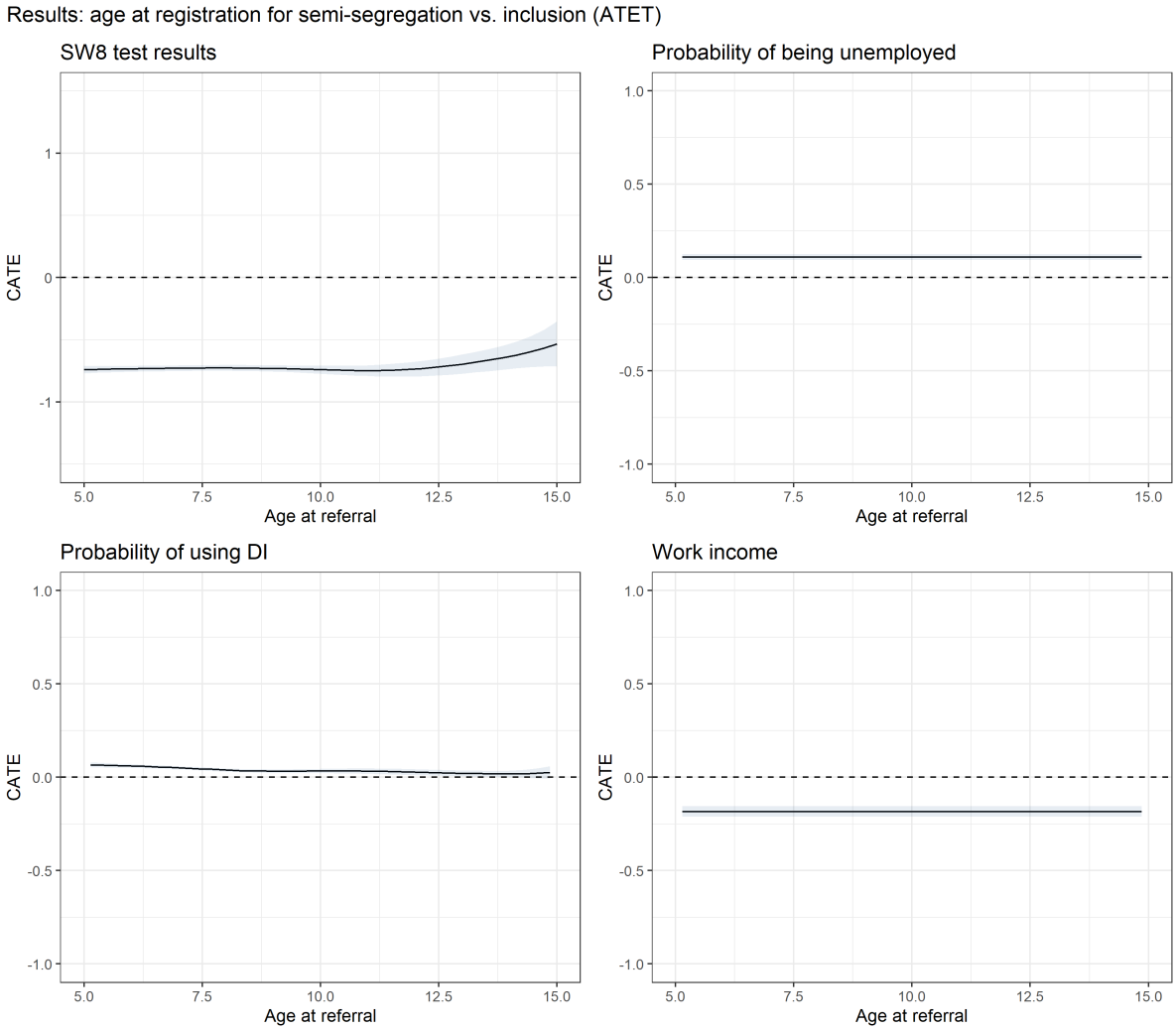


Figure C.4: Heterogeneous ATEs in age at referral to the SPS for "Semi-segregation vs. Inclusion" treatment effects.

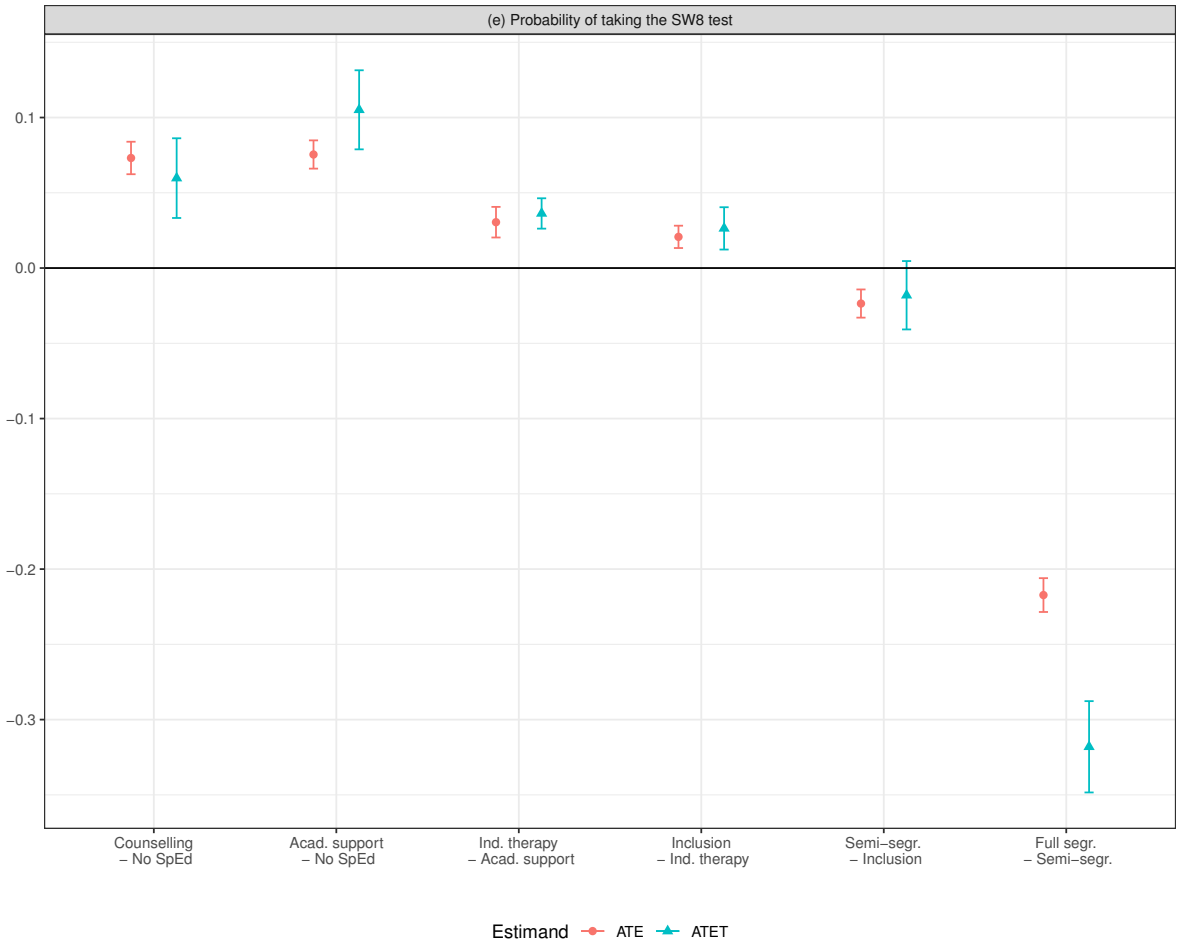
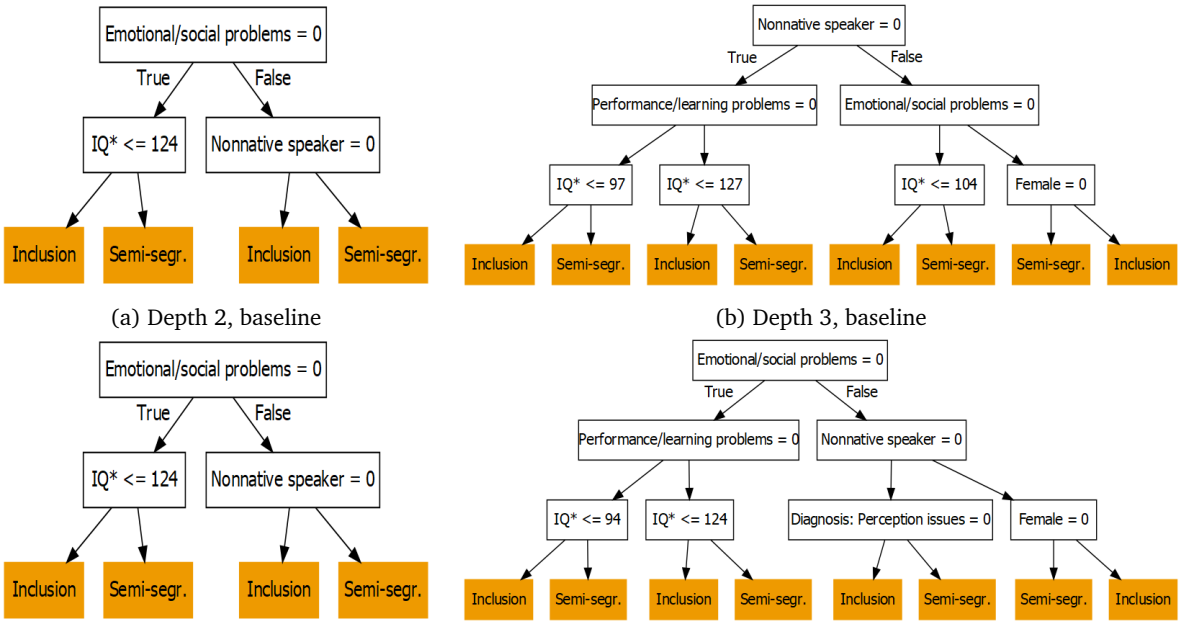


Figure C.5: Pairwise treatment effects for probability of taking the SW8 test

Results for AIPW ATE, ATET and ATO estimations. Common support is enforced. Pairwise treatment effects can be read as follows: "Ind. ther.-None" is the effect of being assigned to individual therapies vs to no therapy, "ISF-None" "KK-ISF" "KK-None" "SON-KK" "SON-None". Nuisance parameters were estimated using an ensemble learner comprising Lasso, Elastic Net and Random Forest with text representations presented in the "data" section. The ensemble weights minimize the out-of-sample MSE as presented in the Appendix. 95% confidence intervals are represented and are based on one sample $t-test$ for the ATE and the ATET and bootstrapping for the ATO. Tables with point estimates are available on request. *Source: SPS*.



(a) Depth 2, baseline

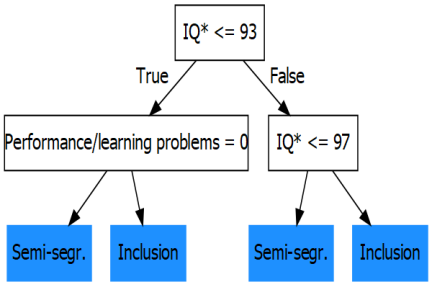
(b) Depth 3, baseline

(c) Depth 2, baseline + diagnosis

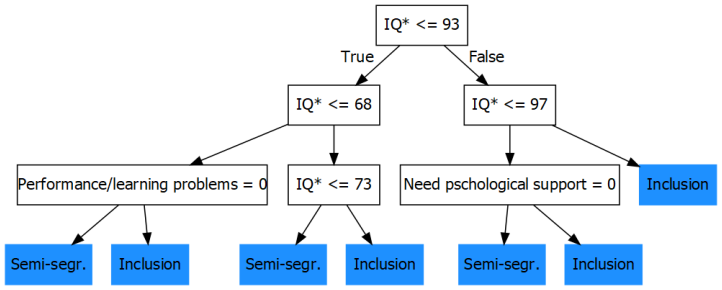
(d) Depth 3, baseline + diagnosis

Figure C.6: Optimal policy trees for test scores

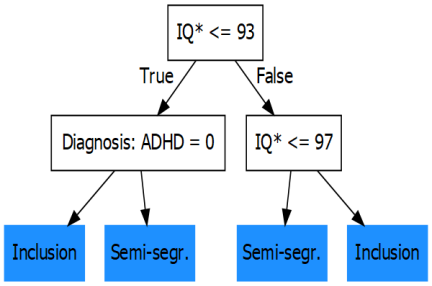
Decision trees for optimal policy allocations of depth 2 and 3 are depicted. "Baseline" means that only baseline students' characteristics are included, and "baseline + diagnosis" includes both baseline characteristics and diagnosis characteristics extracted from psychological records with the dictionary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



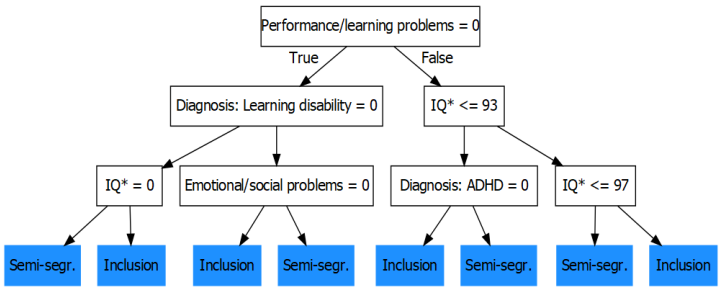
(a) Depth 2, baseline



(b) Depth 3, baseline



(c) Depth 2, baseline + diagnosis



(d) Depth 3, baseline + diagnosis

Figure C.7: Optimal policy trees for probability to be employed

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.