Estimating returns to special education

Combining machine learning and text analysis to address confounding

Aurélien Sallin, University of St. Gallen

Motivation

- **Special education (SpEd) is important**
  - 10% to 25% of students in OECD countries have Special Needs (SEN)
  - Childhood interventions have a long-lasting impact (e.g., Currie, 2001; Heckman et al., 2013; Krueger and Whitmore, 2001)
  - These programs are costly
- **Policy debate: inclusive SpEd is preferred to segregated SpEd**
  - In the US, 30% were mainstreamed in the 1990, 65% in 2017
  - "Political decision" (Lindsay, 2007)
- **SpEd programs are difficult to evaluate empirically**

Research questions

What are returns to SpEd in terms of academic success (test score), and labor market integration (unemployment, disability insurance, wages)?

Today's focus:

Is inclusion more beneficial than semi-segregation for students with SEN?

- What is the average treatment effect of inclusion in comparison to segregation? $\rightarrow$ **ATE**
- Is the effect heterogeneous (individual effects and subgroup effects)? $\rightarrow$ **CATE**
- Can we better allocate students with SEN from segregation to inclusion to maximize academic performance? $\rightarrow$ **Optimal allocation**

Identification and estimation in a nutshell

How do I identify these effects?

- **Selection into programs**: students in inclusion and semi-segregation differ in background characteristics
- I use **written psychological records** for identification of causal effects
  - Use of causal ML, text mining, and NLP to model the treatment assignment process
  - Unprecedented possibility to control for the treatment assignment process.
- **Selection on observables:** cohort, psychologist, and school differences offer variation in treatment assignment

Why is estimation difficult?

- Complex, unstructured, and detailed data: **curse of dimensionality**
  - ML enhances statistical efficiency and improves parametric specifications
  - Double machine learning for flexible program evaluation

Main findings in a nutshell

- **Higher returns from inclusive programs** in comparison to segregated programs
- But **segregation is not worse for all students** with SEN
  - Nonnative speakers and students with social and emotional problems
- Policy rule: **full inclusion** is the policy that maximizes academic performance for students traditionally segregated in small classes
  - lowers overall program costs
  - has negligible effects on mainstreamed students

Contributions of this study (I)

- Short-term (academic) effects of SpEd are well documented, but not long-term effects
  - Positive effects: academic performance (Hanushek et al., 2002; Schwartz et al., 2021) and high-school completion (Ballis and Heath, ming)
  - Zero to negative effects: academic performance (Dempsey et al., 2016; Keslair et al., 2012; Morgan et al., 2010) and high-school completion (Kirjavainen et al., 2016; McGee, 2011)

→ **I assess short- and long-term returns to inclusion and semi-segregation for students with SEN**

- Literature on the effects of inclusion in comparison to segregation for students with SEN is inconclusive
  - Positive effects: (e.g., Cole et al., 2004; Freeman and Alkin, 2000; Peetsma et al., 2001)
    Zero to negative effects: (Lindsay, 2007)

→ **I look at the average and conditional effects of inclusion and segregation**

→ **I conduct optimal policy allocation for inclusive vs. segregated programs**

Contributions of this study (II)

- SpEd has been so far considered as a “single intervention”
  - focus on learning disabilities or students with mild SEN in mainstreamed education (Ballis and Heath, ming; Hanushek et al., 2002; Keslair et al., 2012; Lavy and Schlosser, 2005; Schwartz et al., 2021)
  - without the ability to disentangle between inclusive and segregated programs.

→ **I disentangle the effects of different programs within SpEd**

→ **Today: inclusion and semi-segregation**

- First application of text analysis, NLP, and causal machine learning for program evaluation
  - Text mining and NLP to model the treatment assignment process (e.g., Egami et al., 2018; Gentzkow et al., 2019; Keith et al., 2020; Mozer et al., 2020; Roberts et al., 2020)
  - Causal Machine Learning for treatment evaluation (e.g., Chernozhukov et al., 2018; Knaus, 2021)

Setting

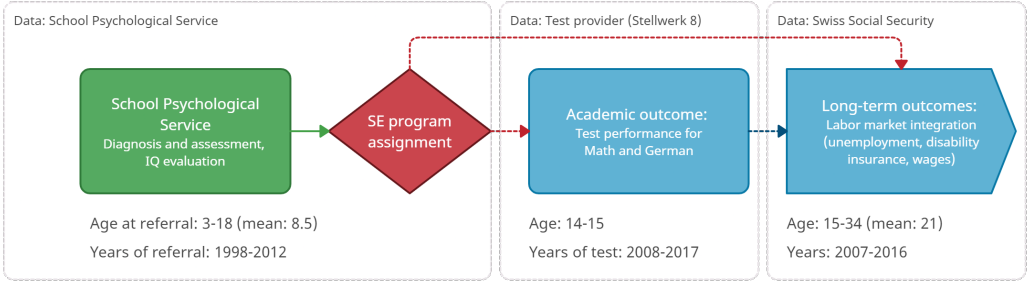
1. **Public school system**: universe of students with SEN from the primary schools of the Canton of St. Gallen
   - School system and SpEd comparable to most OECD countries
   - Around 6% of the Swiss population, both rural and urban

2. **External** and **independent SEN assessment**
   - School psychologists are responsible for diagnoses and intervention decisions (not schools nor parents)

3. **Inclusive SpEd is promoted as a substitute for segregation** in small classes
   - Both target children with behavioral, learning, and school difficulties
   - Political decision to move from semi-segregation to inclusion

Tagblatt article | Placement share segregation-inclusion

Data



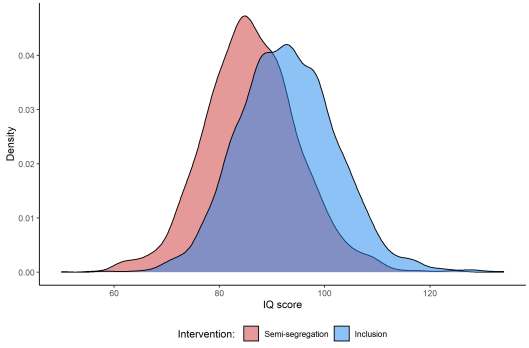
- N = 17,822 students are sent to the SPS in total
- Today, focus on 4,395 students in inclusion and semi-segregation

Text example Summary stats.

Baseline covariates for inclusion and semi-segregation

	Inclusion	Semi-segregation
Female	0.46	0.43
Nonnative	0.11	0.28
Social/emotional problems	0.18	0.22
Performance/learning problems	0.94	0.88
Psychological issues	0.16	0.15
Age referral	8.75	9.11
	(2.03)	(2.50)
Number of visits to SPS	10.3	13.6
	(7.72)	(9.25)

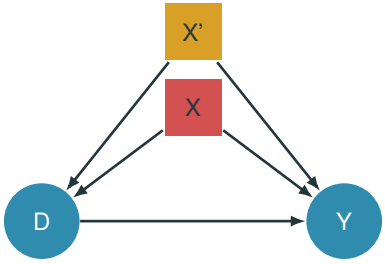
Main covariates



IQ distribution

Differences in schools | Wordcloud

Identification

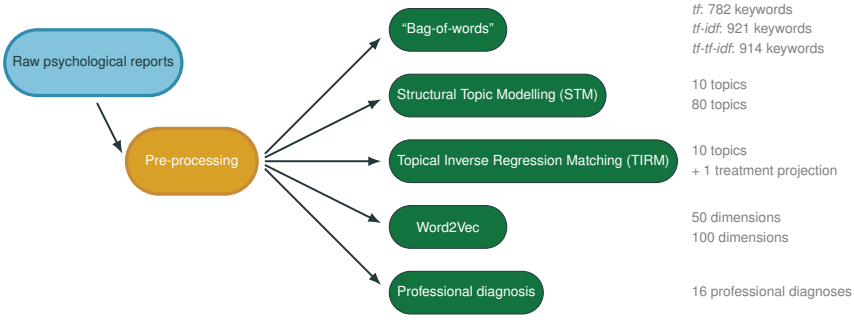


Graph for selection on observables

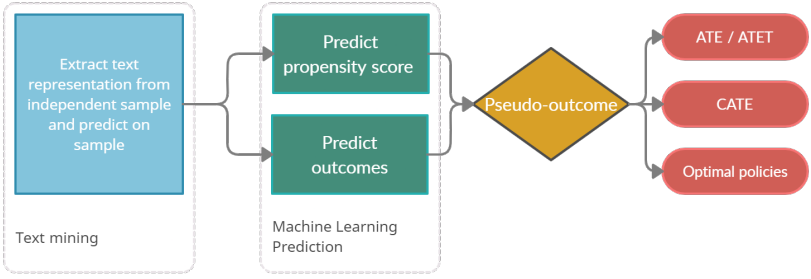
- **Selection-on-observables** setting (e.g., Imbens and Rubin, 2015)
- Conditional Independence Assumption (CIA): available covariates used in the literature (**X**) and placement procedure observed through text (**X'**)
- Sources of variation: cohorts, psychologists, and schools

Identification: Text

- Challenges of using psychological records: **dimensionality reduction** and **comparability**



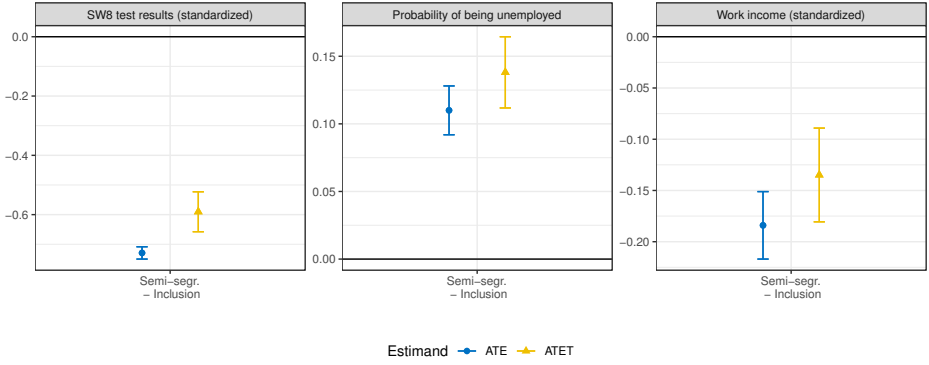
Estimation: Double Machine Learning



- Pseudo-outcome score under each treatment: **doubly-robust AIPW score**
- **Pairwise treatment** effects for each treatment

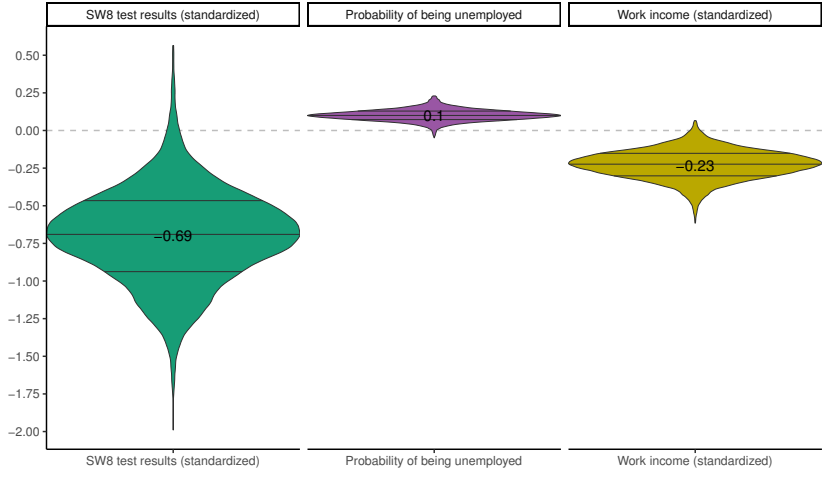
Econometric framework | ML framework

Main results: inclusion vs. semi-segregation



IATE distribution
CATEs along school characteristics

Is inclusion best for all?: Predicted individual effects



Is inclusion best for all?

Individual effects (IATEs):

- Classification analysis 

GATE analysis:

- Group averages over policy-relevant characteristics 

Summary: lowest negative effects of segregation for

- **Academic performance**: nonnatives, children with social/emotional problems, students with psychological problems
- **Unemployment and wages**: nonnatives, and nonnatives with social/emotional problems.


From research to policy...

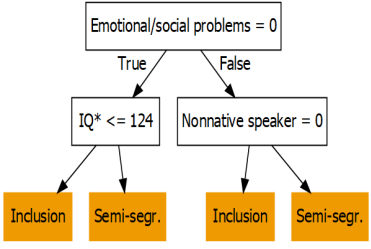
Optimal placement analysis

Could we devise a better placement policy for students in inclusion or semi-segregation?

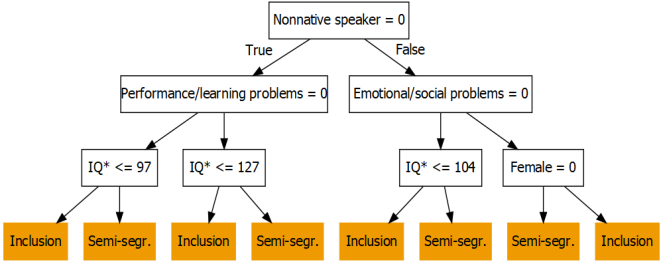
- A **policy** assigns individuals to programs according to a subset of their characteristics
- The **policy value** is the average population outcome under the policy
- The **optimal policy** maximizes the average potential outcome for the population
  - Estimation: “Policy trees” with fixed depth (Athey and Wager, 2021)

- Example:
  - Policy I: assign all SEN students to inclusion
  - Policy II: assign all SEN students to semi-segregation
  - Policy III: assign all nonnative speakers to semi-segregation and all other students to inclusion

Policy allocation: examples of policy rules



(a) Tree of depth 2



(b) Tree of depth 3

Policy allocation: optimal allocation

Test scores

- **Actual allocation**: 69% of students sent to inclusion, 31% to semi-segregation. Costs: 63,925mm CHF.
- **Proposed allocation**: 96% of students sent to inclusion, 4% to semi-segregation. Costs: 60,271mm CHF.

→ **Gain in policy value: 0.17 test score standard deviations. Cost reduction: -6%**

Probability of employment

- **Actual allocation**: 53% of students sent to inclusion, 47% to semi-segregation. Costs: 64,958mm CHF.
- **Proposed allocation**: 83% of students sent to inclusion, 17% to semi-segregation. Costs: 61,007mm CHF.

→ **Gain in policy value: 0.12 percentage points in probability of employment. Cost reduction: -6%**

Table results

Policy allocation: conclusions

Conclusions

- Policy rules improve aggregate academic performance and probability of employment in comparison to the actual assignment
- Important variables for policy assignment are IQ, gender, nonnative status, and emotional/behavioral problems
- **Stability tests**: sending all students to inclusion is almost as good as the proposed optimal policy
- **Negative spillover effects** generated by the reallocation of non-SEN students are minimal

Conclusion

- **Higher returns from inclusive programs** in comparison to segregated programs:
  - Inclusion increases school performance and reduces the probability of unemployment
  - Robust across many robustness checks
- **Better understanding of intervention mechanisms**
  - Effects are heterogeneous across subpopulations of students with SEN
  - Nonnative students and students who are “disruptive” would be as well-off in semi-segregation
- **Policy: reallocation policies could increase average outcomes at lower costs**:
  - Full inclusion appears to be the preferred policy
  - Inclusion incurs only negligible costs on the population of maintreamed students


Thanks for your attention!
asallin.github.io

Newspaper article on the introduction of inclusion measures

TAGBLATT

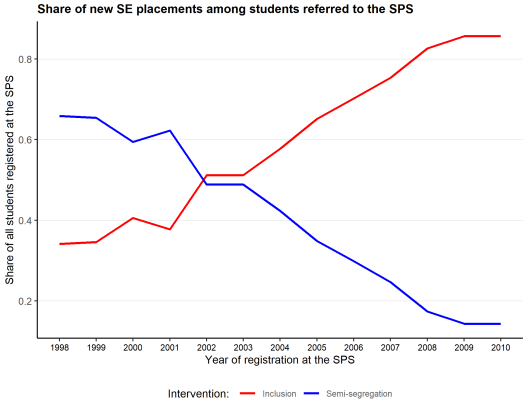
Integrieren statt separieren

Das städtische Förderkonzept wird in den nächsten vier bis sechs Jahren sukzessive umgesetzt. In diesem Schuljahr werden erstmals keine Erstklässler in Kleinklassen eingeteilt. Die Lehrpersonen sollen stärker unterstützt werden.

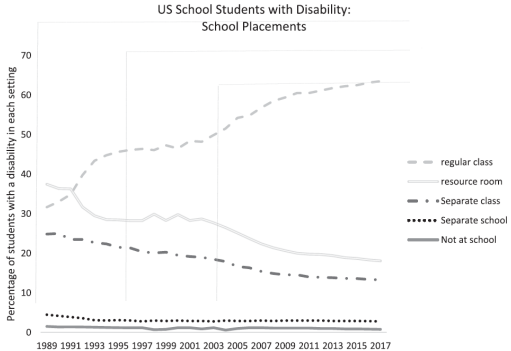
Yvonne Bugmann

09.08.2010, 01.04 Uhr

Placement share: inclusion and semi-segregation

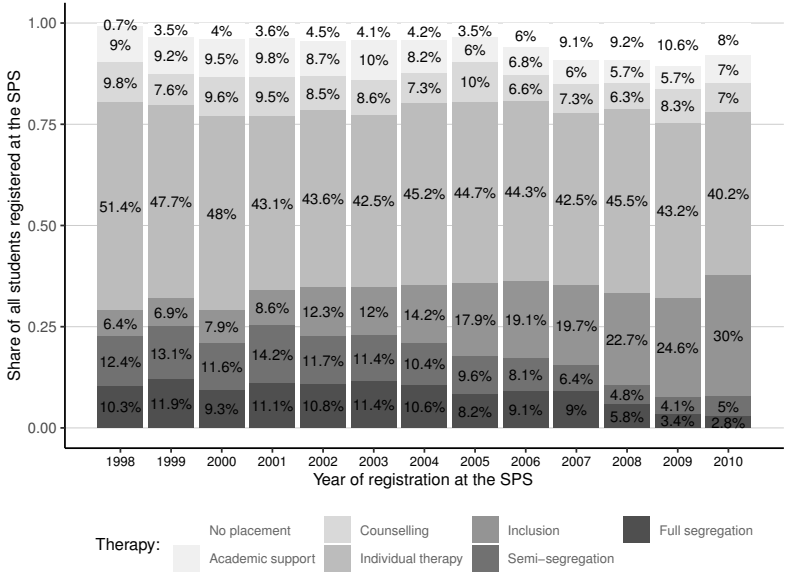


Source: SPS

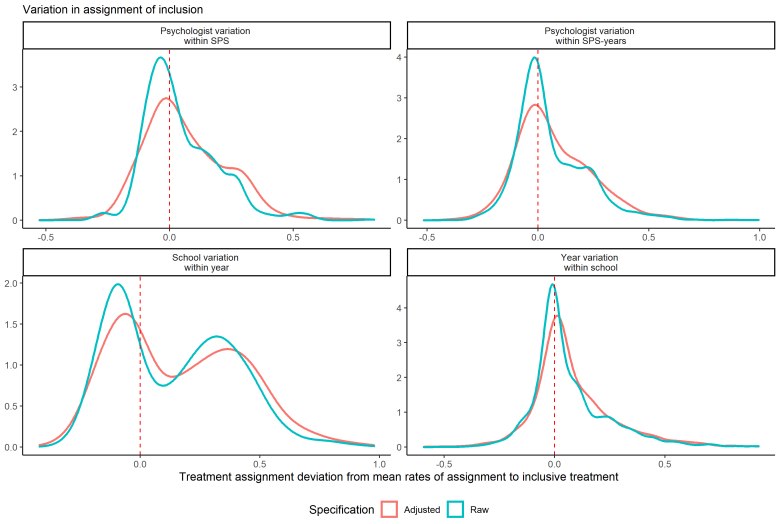


Source: De Bruin (2019)

	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


Share of new SE placements among students referred to the SPS
Share of all students registered at the SPS
1.00
0.75
0.50
0.25
0.00
0.7% 3.5% 4% 3.6% 4.5% 4.1% 4.2% 3.5% 6% 9.1% 9.2% 10.6% 8%
9% 9.2% 9.5% 9.8% 8.7% 10% 8.2% 6% 6.8% 6% 5.7% 5.7% 7%
9.8% 7.6% 9.6% 9.5% 8.5% 8.6% 7.3% 10% 6.6% 7.3% 6.3% 8.3% 7%
51.4% 47.7% 48% 43.1% 43.6% 42.5% 45.2% 44.7% 44.3% 42.5% 45.5% 43.2% 40.2%
6.4% 6.9% 7.9% 8.6% 12.3% 12% 14.2% 17.9% 19.1% 19.7% 22.7% 24.6% 30%
12.4% 13.1% 11.6% 14.2% 11.7% 11.4% 10.4% 9.6% 8.1% 6.4% 4.8% 4.1% 5%
10.3% 11.9% 9.3% 11.1% 10.8% 11.4% 10.6% 8.2% 9.1% 9% 5.8% 3.4% 2.8%
1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010
Year of registration at the SPS
Therapy:
No placement
Academic support
Counselling
Individual therapy
Inclusion
Semi−segregation
Full segregation

Sources of variation in inclusion assignment



Estimation: Double Machine Learning

The estimands of interest are identified by the **doubly-robust AIPW score** for the ATE and ATET

(e.g., Chernozhukov et al., 2018; Li and Li, 2019; Robins et al., 1994)

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

And the following aggregations:

$$\begin{aligned}
\mathsf{APO}_d &= E[\Gamma^h(d, X_i)] \\
\mathsf{ATE}_{d,d'} &= E[\Gamma^h(d, X_i) - \Gamma^h(d', X_i)] \\
\mathsf{ATET}_{d,d'} &= E[\Gamma^h(d, X_i) - \Gamma^h(d', X_i)|D_i = d] \\
\mathsf{GATE}_{d,d',z} &= E[\Gamma^h(d, X_i) - \Gamma^h(d', X_i)|Z_i = z]
\end{aligned}$$

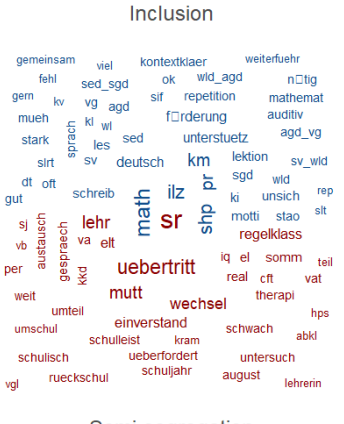
For the ATE:

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

School covariates for inclusion and semi-segregation

	Inclusion	Semi-segregation	Stat. difference
Differences in schools characteristics			
School: percent nonnatives	0.151	0.261	***
School: percent SEN students	0.200	0.157	***
School: social index	0.930	1.007	***
School: total number of students	84.312	170.838	***
School: expenditures per student (2017)	0.055	-0.374	***
School: urban	0.169	0.542	***

Wordcloud for inclusion and semi-segregation



Notes: this boxplot shows the words that are the most distinctive of both inclusive and segregated programs. Words that would overlap and be frequent in both groups are removed.

Identifying assumptions

Assumptions necessary for **causality** (e.g., Imbens and Rubin, 2015):

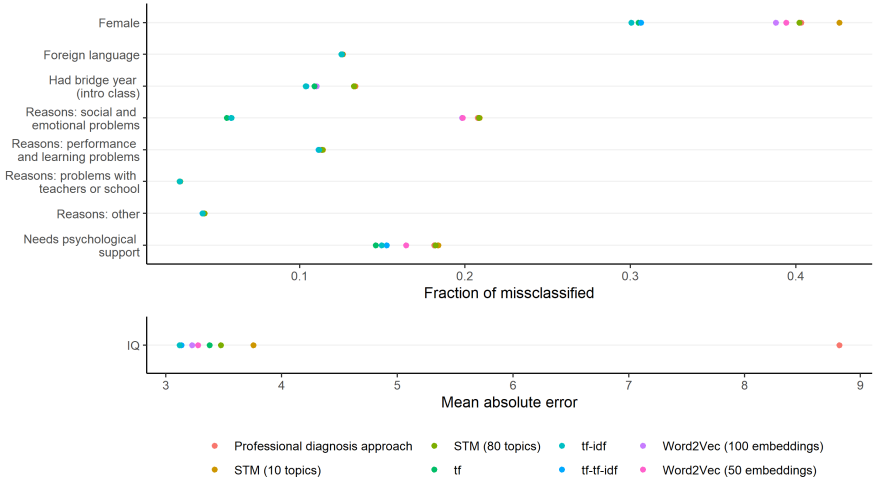
- **Conditional independence assumption (CIA)**:
  - Text!
- **Common support**:
  - Variation of programs within SPS jurisdictions (psychologist variations), across years, and across schools
- **Exogeneity of confounders**:
  - Only pretreatment covariates are used ($\rightarrow$ Text is extracted before the first treatment assignment)
  - Covariates are not a function of treatment: caseworkers are not perfectly able to assign treatment based on expected returns, as there is variation in program assignments across schools and across years.
- **"No interference" (SUTVA)**:
  - Spillovers (peer) effects are "part" of each program
  - Budgeting of SpEd is centralized

Example of text: fake comment

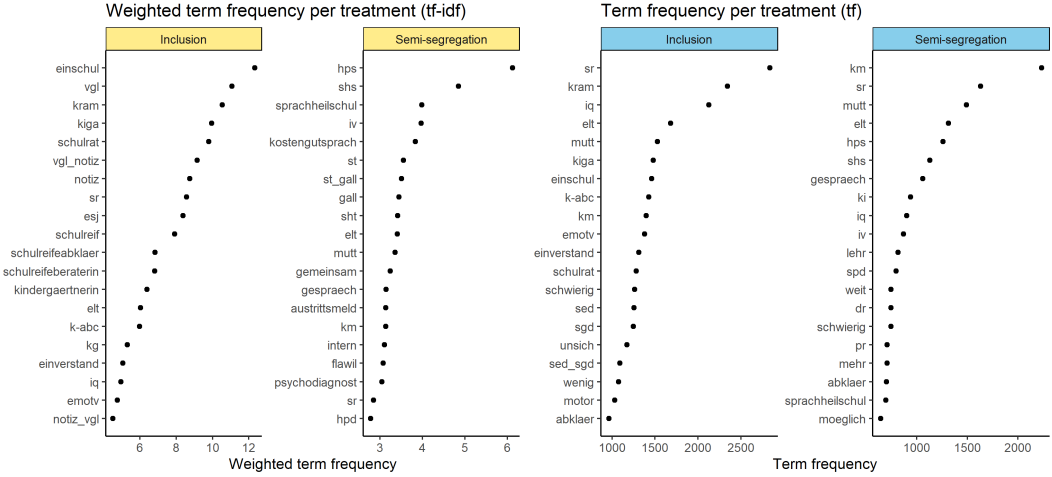
The mother contacted us because MM does not feel good in school, does not get along with the teacher, feels that he is treated unfairly. Child feels that he is not well accepted at school. I administered an IQ test: HAWIK-III IQ = 83 (VT: 92 / HT: 78). The child has difficulties with a general overview and the organization of his work, he is very talkative, and also marginal to failing school performance. MM has high expectations on himself and does not handle criticism well. Is able, with some help, to see that he also contributes to the situation. The mother talked to the teacher, which did not solve anything, but she will try again. The teacher only blames the child. Longer discussion with the mother, she feels like she is not taken seriously by the school. She welcomes a discussion between MM, the teacher and myself. DATE-24.05.1970 Preparation for the discussion: [the teacher] experiences MM as ''spoiled'' and influenced by his mother. The teacher and MM are able to agree on some points that the teacher expects, and the teacher is ready to give MM positive feedback.

DATE-28.02.1980 Phone session with the mother regarding mobbing, and schedule of an appointment DATE-11.03.1980 Cancellation of the appointment. The kid does not want it, the mom will contact us if needed. Discussion will be held with or without the child.

Identification: Text

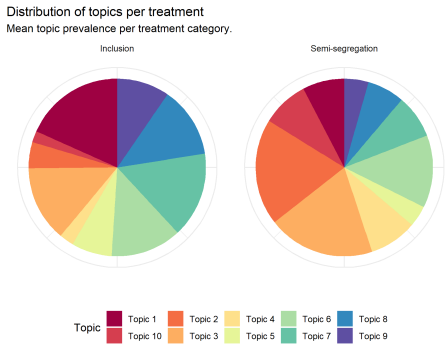


Learning from text: tf and $tf - idf$



Learning from text: stm

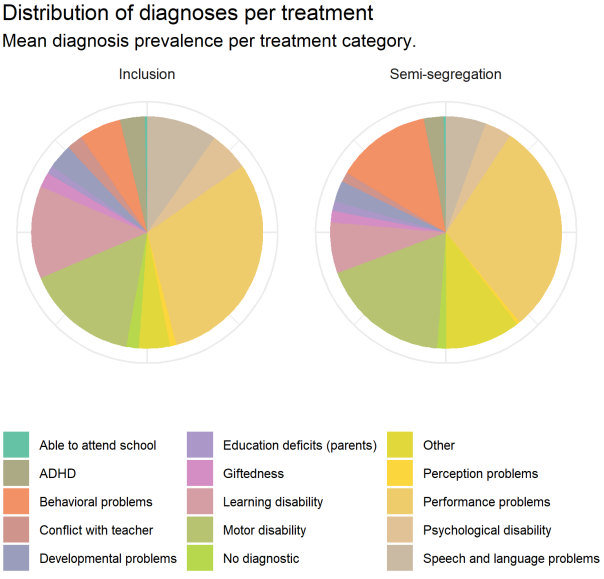
Topic	Highest probability	Most frequent and exclusive
1	sr, mutt, km, schulrat, les, pr, kiga, vgl, lehr, pr	sr_lekt, vb, jug, u'ergebnis, iq_sed, gemeinsam_auswertungsgespraech, trog-d, vgl_notiz, proz, sht
2	iq, elt, ki, besprech, rechn, hawik, einschul, notiz, ke, kl	untersuch_hawik, herrn, inform, 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, legasthenieschlussbericht, diagnost_termin, wn
8	uebertritt, situation, noetig, therapi, mueh, sv_wld, einverstand, austausch, termin, pl	schlussgespraech_sr, sond, thera, schulpsycholog_abklaer, lektion_woechent, psychodiagnost, 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



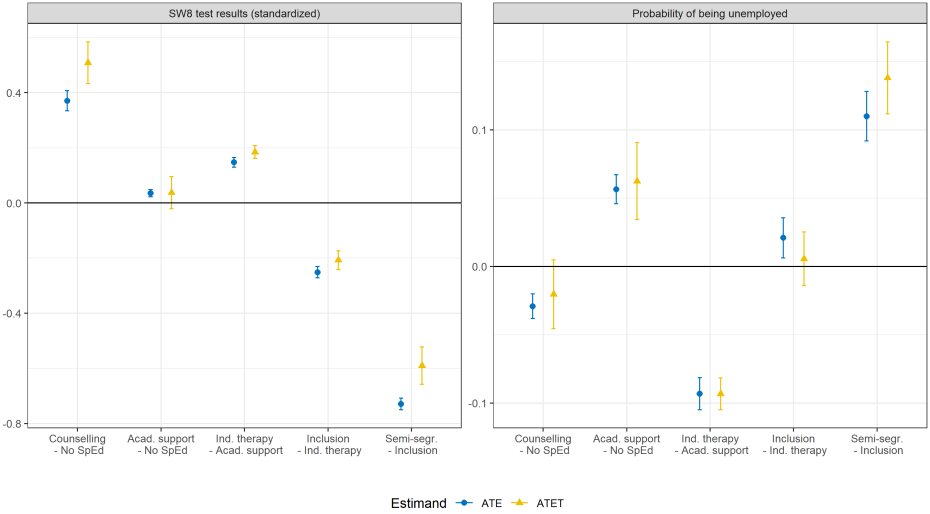
Learning from text: Word2Vec

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

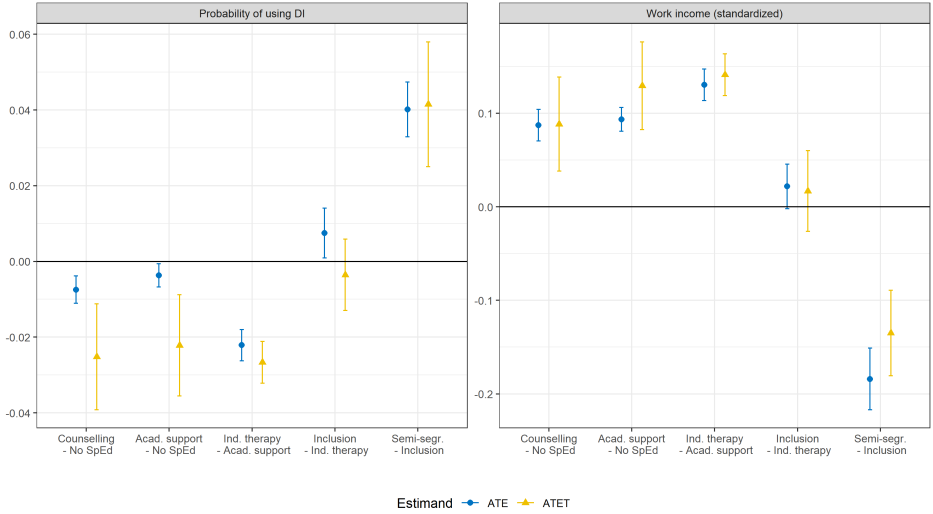
Learning from text: predicted diagnoses



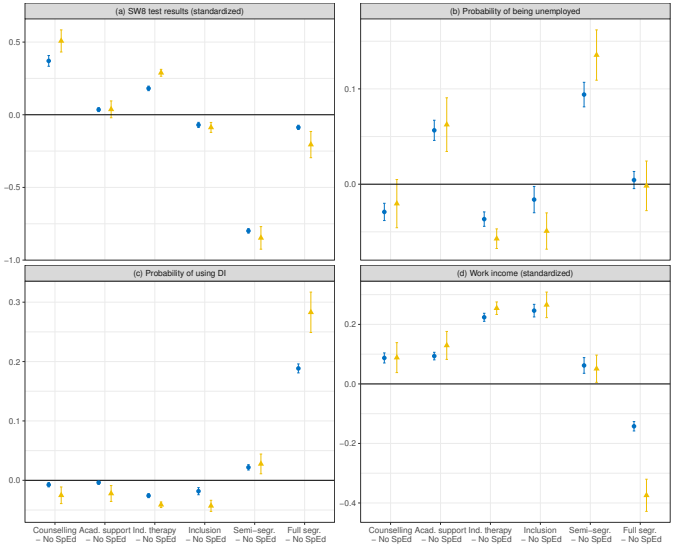
Main results: pairwise comparisons I



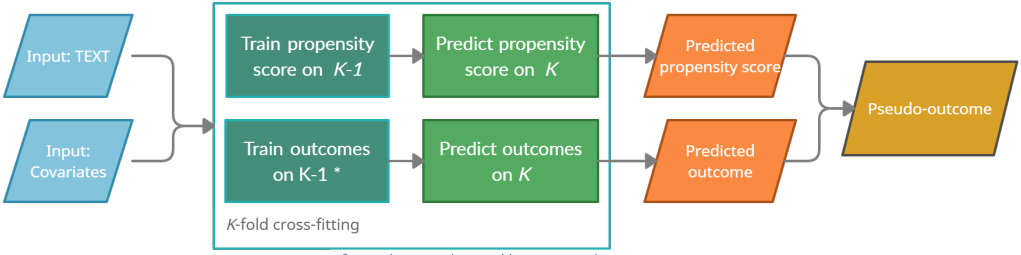
Main results: pairwise comparisons II



Main results: comparison with no treatment

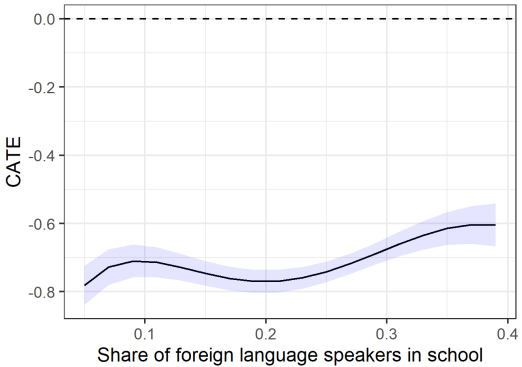


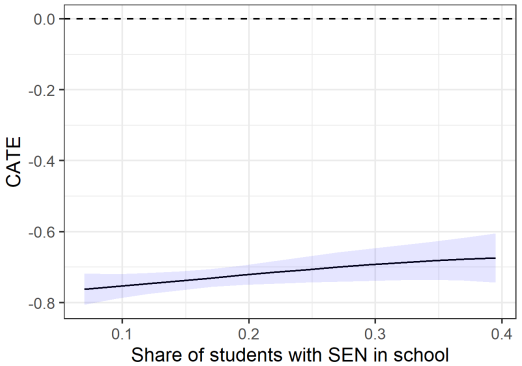
Estimation: ML algorithms



- Estimation of two **nuisance parameters**: $\mu(d,x) = E[Y_i|D_i = d, X_i = x]$ and $p_d(x) = P[D_i = d|X_i = x]$
  - Lasso and binomial lasso, Elastic Net and binomial Elastic Net, Random Forest, (...)
  - Outcomes models under treatment d are trained on the subgroups observed in d
- **Ensemble learner**: weighted predictor such that RMSE per fold is minimized
  - $9 \times 3 \times 7 = 189$ estimations for the propensity score, $9 \times 3 \times 7 \times 4 = 756$ estimations for outcome...
- K**-fold cross fitting** is crucial for statistical properties of the AIPW score
  - prevents the correlation of the model parameters estimation step, and the treatment effect estimation step (overfitting bias)

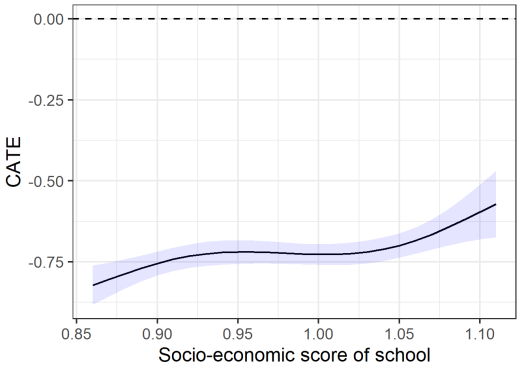
CATEs for test scores along school characteristics I

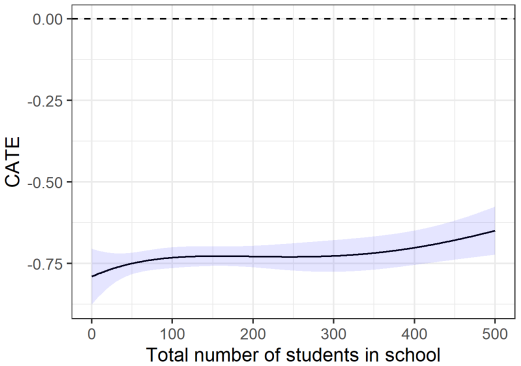




Notes: The y axis is the pseudo-outcome for the difference between inclusion and semi-segregation in test scores. CATEs are nonparametrically estimated following (Fan et al., 2020; Zimmert and Lechner, 2019). Kernel regression on the pseudo-outcome with 0.9 smoothing. 95% confidence intervals are represented.

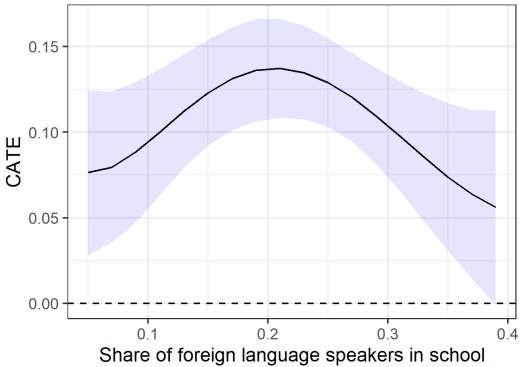
CATEs for test scores along school characteristics II

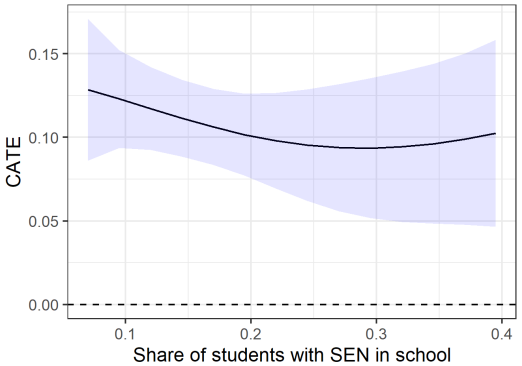




Notes: The y axis is the pseudo-outcome for the difference between inclusion and semi-segregation in test scores. CATEs are nonparametrically estimated following (Fan et al., 2020; Zimmert and Lechner, 2019). Kernel regression on the pseudo-outcome with 0.9 smoothing. 95% confidence intervals are represented.

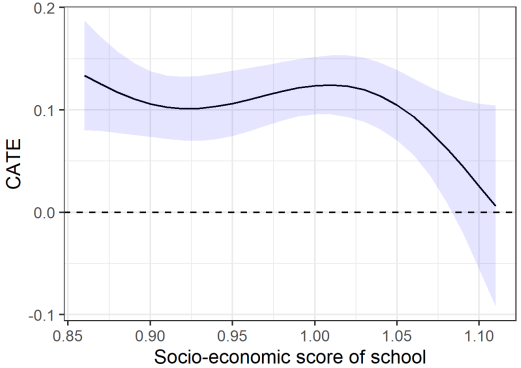
CATEs for test scores along unemployment probability I

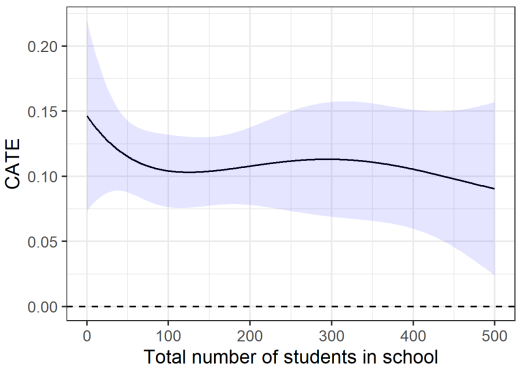




Notes: The y axis is the pseudo-outcome for the difference between inclusion and semi-segregation in unemployment probability. CATEs are nonparametrically estimated following (Fan et al., 2020; Zimmert and Lechner, 2019). Kernel regression on the pseudo-outcome with 0.9 smoothing. 95% confidence intervals are represented.

CATEs for test scores along unemployment probability II



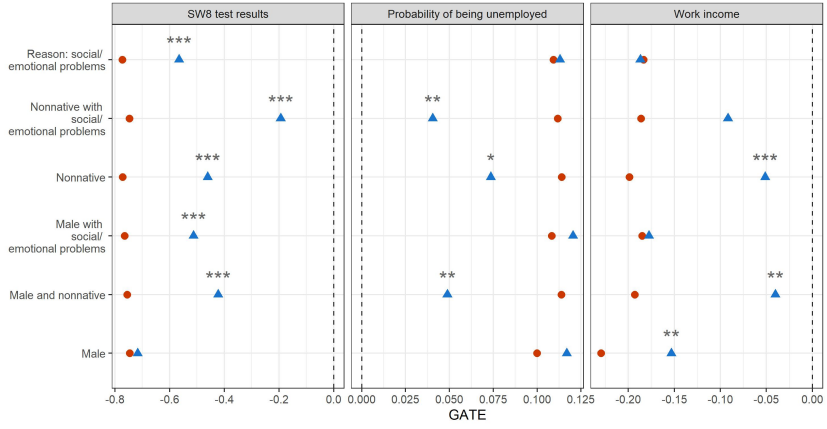


Notes: The y axis is the pseudo-outcome for the difference between inclusion and semi-segregation in unemployment probability. CATEs are nonparametrically estimated following (Fan et al., 2020; Zimmert and Lechner, 2019). Kernel regression on the pseudo-outcome with 0.9 smoothing. 95% confidence intervals are represented.

Inclusion vs. semi-segregation: IATEs

	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 $\times$ Female	0.04	0.13	**0.339**	0.03	0.08	**0.220**	0.02	0.08	**0.284**
Nonnative $\times$ Social/emotional	0.01	0.11	**0.433**	0.01	0.06	**0.309**	0.00	0.06	**0.333**

Inclusion vs. semi-segregation: GATEs



Policy allocation: optimal allocation

	% Students sent to inclusion	% Students sent to semi-segregation	Policy effect	Costs per year (in mm CHF)	Percent of actual costs
Test scores (N = 2988)					
Actual allocation	0.69	0.31	**0**	63,925	1
Depth 2 and baseline variables	0.96	0.04	0.17	60,271	0.94
Depth 3 and baseline variables	0.96	0.04	0.19	60,235	0.94
Depth 2 and diagnosis variables	0.96	0.04	0.17	59,762	0.93
Depth 3 and diagnosis variables	0.97	0.03	0.19	60,145	0.94
Probability of employment (N = 2939)					
Actual allocation	0.53	0.47	**0**	64,958	1
Depth 2 and baseline variables	0.83	0.17	0.12	61,007	0.94
Depth 3 and baseline variables	0.87	0.13	0.13	60,490	0.93
Depth 2 and diagnosis variables	0.87	0.13	0.12	60,485	0.93
Depth 3 and diagnosis variables	0.85	0.15	0.14	60,796	0.94

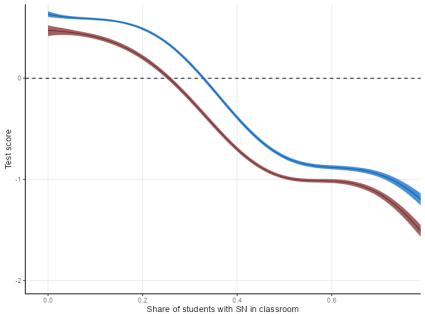
Cost estimates: inclusion (20,000 CHF per student per year); semi-segregation (24,500 CHF per student per year).
Baseline policy value: -0.46 for test scores, 0.67 for probability of employment.

Policy allocation: stability tests

	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.

What about students without SEN?



- **25% *mainstreamed* students have SEN**
  - Random allocation to classroom (see Balestra et al., ming)
- Reallocation of 807 students: 0.3 additional students with SEN per classroom
- **Average Reallocation Effects** (ARE): **-0.0227** test score standard deviations.
- **Conditional ARE**: loss of 0.03 (0.04) test score standard deviations for mainstreamed students without (with) SEN

Robustness checks / further work

Attrition

- Test taking
- Cohort in both datasets

Overlap and large IPW weights

- Average Treatment effect on the population of Overlap (ATO) (Li and Li, 2019)
- Trimming and common support sensitivity analysis

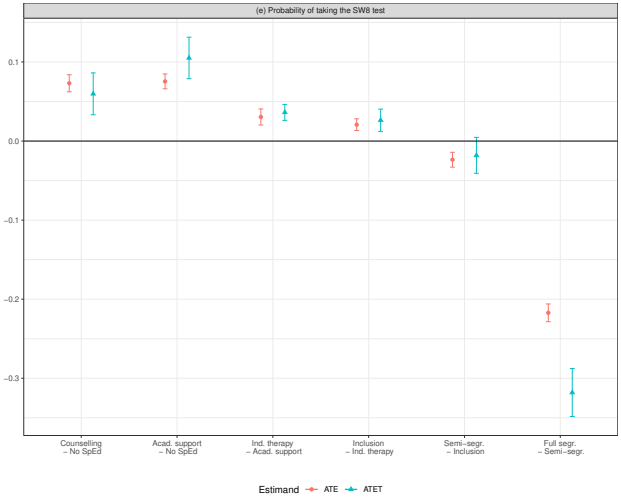
Text

Results without text covariates

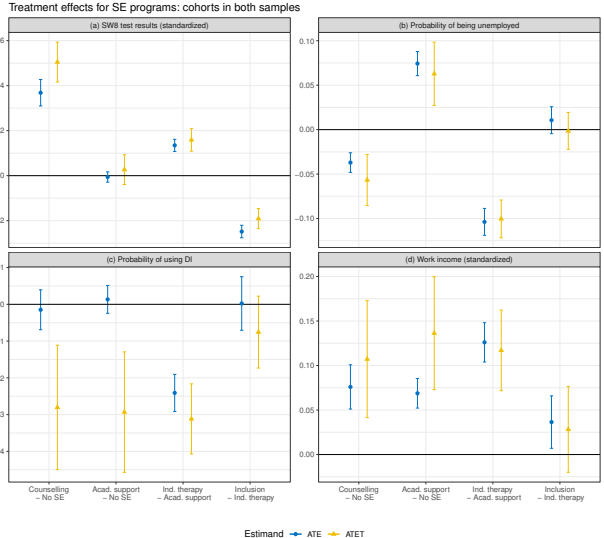
Identification

IV with municipal variation in supply for inclusive SE programs

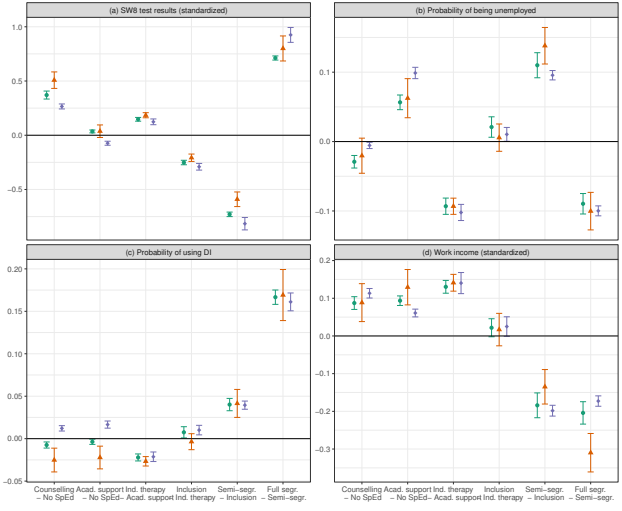
Attrition: test taking



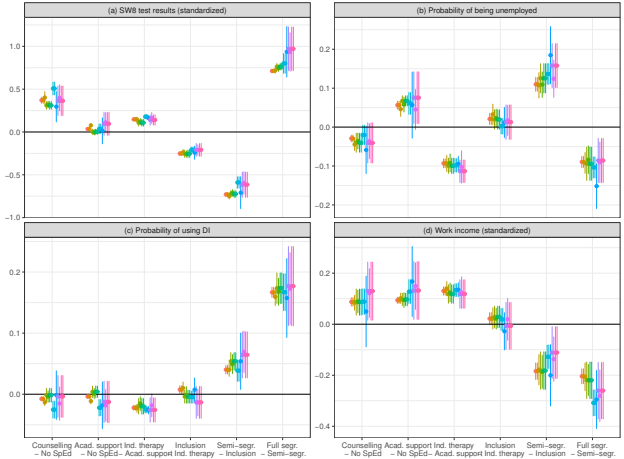
Attrition: cohorts



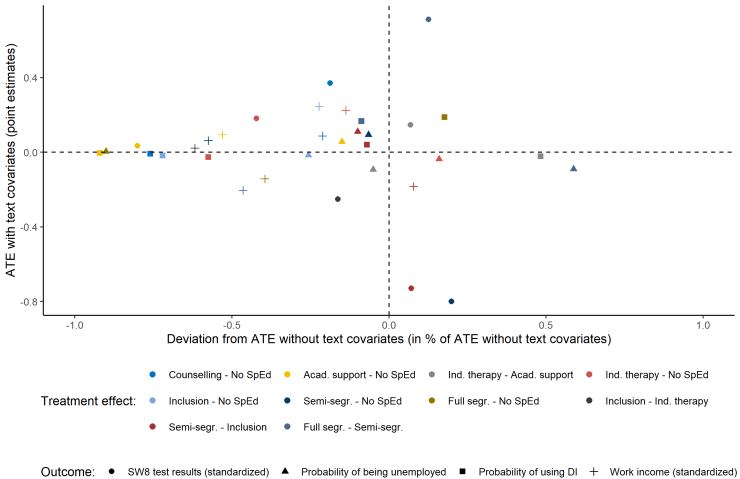
Results of the ATO



Main results with trimming



Estimates without text covariates



References i

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Balestra, S., Eugster, B., and Liebert, H. (forthcoming). Peers with special needs: effects and policies. *Review of Economics and Statistics*.

Ballis, B. and Heath, K. (forthcoming). The long-run impacts of special education. *American Economic Journal: Economic Policy*, 43.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68.

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

Currie, J. (2001). Early childhood education programs. *Journal of Economic Perspectives*, 15(2):213–238.

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

Dempsey, I., Valentine, M., and Colyvas, K. (2016). The effects of special education support on young australian school students. *International Journal of Disability, Development and Education*, 63(3):271–292.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., and Stewart, B. M. (2018). How to make causal inferences using texts. *ArXiv Working Paper*, (1802.02163).

Fan, Q., Hsu, Y.-C., Lieli, R. P., and Zhang, Y. (2020). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, pages 1–15.

Freeman, S. F. and Alkin, M. C. (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, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–74.

References ii

Hanushek, E. A., Kain, J. F., and Rivkin, S. G. (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.

Heckman, J., Pinto, R., and Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, 103(6):2052–86.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Keith, K. A., Jensen, D., and O'Connor, B. (2020). Text and causal inference: A review of using text to remove confounding from causal estimates. *arXiv preprint arXiv:2005.00649*.

Keslair, F., Maurin, E., and McNally, S. (2012). Every child matters? an evaluation of "special educational needs? programmes in england. *Economics of Education Review*, 31(6):932– 948.

Kirjavainen, T., Pulkkinen, J., and Jahnukainen, M. (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.

Knaus, M. (2021). Double machine learning based program evaluation under unconfoundedness. *Working Paper*.

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

Lavy, V. and Schlosser, A. (2005). Targeted remedial education for underperforming teenagers: Costs and benefits. *Journal of Labor Economics*, 23(4):839–874.

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

Lindsay, G. (2007). Educational psychology and the effectiveness of inclusive education/mainstreaming. *British journal of educational psychology*, 77(1):1–24.

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

References iii

Morgan, P. L., Frisco, M. L., Farkas, G., and Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education*, 43(4):236–254.

Mozer, R., Miratrix, L., Kaufman, A. R., and Jason Anastasopoulos, L. (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, T., Vergeer, M., Roeleveld, J., and Karsten, S. (2001). Inclusion in education: Comparing pupils' development in special and regular education. *Educational Review*, 53(2):125–135.

Roberts, M. E., Stewart, B. M., and Nielsen, R. A. (2020). Adjusting for confounding with text matching. *American Journal of Political Science*.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Schwartz, A. E., Hopkins, B. G., and Stiefel, L. (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.

Zimmert, M. and Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv preprint arXiv:1908.08779*.