Class composition, special needs students, and peers' achievement

Simone Balestra [*], [†] Beatrix Eugster [*], [‡] Helge Liebert [*], [§]

December 19, 2016

First draft. Preliminary and incomplete. Comments are welcome.

Abstract

This paper evaluates the effects of classroom composition and inclusive schooling on student achievement. Combining population data on student achievement with psychological examination records, we find that higher levels of special needs students in a class lower students' performance in a standardized test. There is distinct effect heterogeneity: Special needs students themselves and students at the lower end of the achievement distribution suffer the most from higher inclusion. Furthermore, the effect is largely driven by students at the upper end of the special needs distribution, i.e. those with more severe needs and conspicuous behavior. In addition, we find evidence that effects only emerge after the number of special needs students exceeds 15% of students in a classroom.

Keywords: class composition, special needs, peer effects
JEL classification: I21, I28, J14

[*]CDI-HSG, University of St. Gallen, St. Gallen, Switzerland. We are grateful for support from Hermann Blöchlinger, Christian Brühwiler, Claudia Coray, Markus Hartmeier and Ralph Wettach.

[†]E-mail: simone.balestra@unisg.ch.
[‡]E-mail: beatrix.eugster@unisg.ch.
[§]E-mail: helge.liebert@unisg.ch.

1 Introduction

Following the UN Convention on the Rights of the Child in 1989, countries worldwide have adopted more inclusive education practices for children with special needs. Special needs (SN) refer to the requirement of assistance for medical, mental or psychological disabilities. In the United States, the Individuals with Disabilities Education Act (IDEA) mandates that children with special educational needs have the opportunity to be educated in the *least restrictive environment*. About seven million students are eligible for special needs assistance for education under IDEA (U.S. Department of Education 2016). The idea of *mainstreaming*, i.e. schooling SN children in regular education among children without SN, is pervasive in Europe as well (e.g. European Agency for Development in Special Needs Education 2010). In Switzerland for example, inclusive schooling has become the norm for children with special needs, except for those with severe impairments. About a quarter of all primary school students are in contact with special needs services at least once.

In spite of the policy trend towards inclusive education, little is known about how inclusion affects achievement of students. Nonetheless, understanding whether differences in classroom composition generate negative externalities is crucial. Given the public goods nature of classroom education production, disruptive influences by special needs students can potentially impede peers' learning (Lazear 2001).

To fill this gap, this paper evaluates the impact of differential classroom composition in an inclusive education setting on students' educational achievement. We merge psychological health examination records with comprehensive data on standardized tests for the student population of a large Swiss region. These unique data allow us to examine in detail classroom externalities generated by the inclusion of SN students. For identification, we exploit natural variation in the level of SN students per class within schools.

We find that higher levels of SN students in a class are detrimental to their peers' educational outcomes. One additional SN child in a class reduces math test scores by 2.2% of standard deviation. However, there is substantial effect heterogeneity. The effect is mostly driven by children with severe special needs, i.e. those at the upper end of the SN intensity distribution. In addition, it is students at the lower end of the achievement distribution and special needs students themselves that are most negatively affected. Threshold effects are relevant: Negative spillovers occur only after SN students exceed 15% of students in a class.

Our results relate to established findings in the empirical literature on classroom education production and peer effects. A variety of studies evaluate the impact of potentially disruptive children on their peer group's educational outcomes (e.g., Carrell and Hoekstra 2010, Figlio 2007, Kristoffersen et al. 2015). These studies typically proxy disruptiveness by considering parents income, criminal background, children's gender or similar approximate measures. They find detrimental effects of potentially disruptive children on peers' achievement. Related to this, other papers on peer effects in education have established a negative association between peer quality and educational achievement (Burke and Sass 2013, Sacerdote 2001). Our results also relate to the literature on special education. Evaluating special education measures, Hanushek et al. (2002) provide some evidence that regular students are unaffected by the presence of special education students and mainstreaming.

We extend previous findings in several dimensions. First, by using special needs examination records, we focus on a student group of specific policy relevance without relying on imprecise proxies of peer quality or disruptiveness. Second, we study differential levels of SN class composition as opposed to the arrival of a single individual to a school class or even cohort. The variation in class composition in combination with SN intensity measures from the examination data allows us to examine effect heterogeneity in unprecedented detail. Taking the inclusive schooling structure as given, we aim to answer the question how to compose classes optimally to minimize potential negative spillovers. Third, due to data limitations, much of the previous literature defines the school cohort as the relevant peer network and is agnostic towards the shared class structure. We focus on peer effects in classroom education production and explicitly compare results for different definitions of peer environments. Theory posits that the classroom is the relevant peer environment. We show that approximating class externalities at the school cohort level captures peer effects that are qualitatively similar but smaller.

The paper proceeds as follows: Section 2 introduces the institutional setting, section 3 presents the data, section 4 discusses the empirical strategy and section 5 the results. Section 6 concludes.

2 Institutional Background

2.1 Education system

In Switzerland, responsibility for the education system lies with the states (cantons). The majority of schools are public, about 5% of schools are private. Kindergarten and primary school are typically organized at the municipal level. Large municipalities may host several primary schools, few very small municipalities send their children to neighboring municipalities' schools. Secondary schools are organized separately from primary schools and have a larger catchment area. Assignment to schools is based on district residence. The assignment is strict, parents have no influence on the decision, other than moving permanently to another district. Appeals challenging the assignment decision are typically dismissed and rulings regarding exceptions from residence-based assignment are rarely granted (e.g. St. Galler Tagblatt 2016).

Our analysis considers the universe of schools from the canton of St. Gallen with a population of approximately 500,000, the fifth largest state in Switzerland. Schooling is compulsory for eleven years in St. Gallen. Children enter kindergarten in fall of the year in which they have reached age four before August for a duration of two years. After kindergarten, children attend primary school for six years and then transfer to lower secondary school for another three years. Classes remain unchanged within each stage, but are reshuffled when transitioning between stages. At the end of eighth grade, i.e., the second year of secondary school, all children take a mandatory standardized test in Math, German, English and Science.

Tracking occurs after finishing primary school. There are two main tracks, secondary schools either belong to an upper tier (*Sekundarschule*) or a lower tier (*Realschule*). Assignment to either track is based on the primary school teacher's recommendation and the students' performance. Both tracks together account for 95% of students in a cohort. The remaining 5% are enrolled in special education schools or in a rare direct academic track (*Untergymnasium*). While the assignment to secondary school tracks is based on performance, there is no systematic assignment to classes within each school and track. When classes are composed at the beginning of secondary school, administrative staff have no prior knowledge of students. Anecdotal evidence suggests that schools deliberately try to mix their intake cohort based on observable characteristics, i.e. the attended primary school, residence within the catchment area and gender. In the school year 2016/2017, 53,264 children were enrolled in schools in St. Gallen; 10,098 children were enrolled in kindergarten and 27,944 children in primary school, distributed over 83 different school districts. 13,287 children attended one of the 57 secondary schools.

After finishing compulsory education, students either continue secondary education and follow an academic track or apply for an apprenticeship. Switzerland has a strong tradition of vocational education, reflected in a low proportion of students choosing the academic track. In St. Gallen, approximately 14% of students per cohort complete the academic track, allowing them to pursue post-secondary education; compared to a national average of 20%.

Inclusion is an important policy target of public schools. Children of different gender, race, social and economic background as well as children with disabilities should be educated in an inclusive setting whenever possible. Only 3.5% of all schools in Switzerland are segregated special education schools for children with severe disabilities, learning problems or conspicuous behavior. While historically it was also common to school SN children within small segregated classes embedded in public schools, this form of segregation became less widespread during the past decade. During the school year 2016/2017, 1,112 children were enrolled in small SN classes in the canton of St. Gallen.

To take into account specific needs of SN students, the standard school curriculum can be individually adapted to a certain degree. Vulnerable children or those with developmental delays are screened at school entry, and can be offered the possibility of a transition year between kindergarten and primary school to foster school readiness. In the school year 2016/2017, 558 children were enrolled in such a transition year. Later, students with strong special needs might be accompanied by a SN teacher full time or during specific lessons. Because such measures are costly, they are only implemented upon recommendation of the School Psychological Service following an examination. These inclusive measures are often complemented by individually assigned therapies, e.g., speech, dyslexia/dyscalculia or behavioral therapy. In general, interventions are predominantly taken during primary school, when children are still in a formative age. Even though children may still have SN and school problems, support during secondary school is considerably reduced. Assistance in secondary school mainly consists of additional tutoring, which is provided to less than 1.5% of students.

2.2 The school psychological service

The school psychological service (SPS) is a centralized service provider for all schools in the canton of St. Gallen. It is split into two separate administrative units for the main city St. Gallen and the remainder of the canton, which is served by seven regional offices. The SPS provides diagnosis and counseling for children, parents and teachers for school-related problems. This includes diverse issues such as general counseling, diagnosis of learning disabilities and correspondence with physicians, monitoring developmental deficiencies, conflict mediation and developing schooling strategies for children with severe conspicuous behavior.

In most cases, the teacher registers a child for counseling with the SPS. In a few cases, requests for counseling are filed directly by a child's parents. A counseling unit typically begins with a diagnosis session which may include several diagnostic tests, followed by discussions with the relevant stakeholders, and finishes with a recommendation for further action, e.g. delaying school entry, additional coaching or therapy. Schools have some freedom in implementing the recommendation, but rarely deviate. Because of the diverse issues the SPS treats, about one fourth of all children are in contact with it at least once. From 2004 until 2015, on average about 3,300 children where registered with the SPS per year. The average age at first registration is eight years, typically when children are in the third year of primary school.

3 Data

Data for this study stem from two sources. We use data on achievement test scores and individual characteristics for the population of students enrolled in eighth grade in the canton of St. Gallen during the years 2008 to 2016. We then merge these data with administrative records from the SPS.

Our outcome data on achievement test scores are based on a compulsory standardized test taken in eighth grade (*Stellwerk8*) and are available for the subjects Math, German, English, and Science. Our main specifications focus on the test scores in Math and German, since these two subjects are compulsory for all tracks. In additional models we also look at test scores in English and Sciences. English, while not strictly compulsory, is chosen by 99% of students as an elective. We define Sciences as the average score across the subjects biology, chemistry, and physics. These are mandatory electives of which students typically have to choose at least two. Other elective subjects, e.g. French or Latin, are tested less often and scores are not available for all students. The test is administered between February and April of the second year of secondary school, towards the end of the school year.

Stellwerk8 is a norm-referenced, self-scoring, adaptive, computer-based exam similar in spirit to the Graduate Record Examination (GRE). All students in grade eight except those enrolled in special education schools are tested. Students do not necessarily face the same set of questions. The software chooses questions such that they correspond to an ability measure computed based on the answers to all previous questions. The final test score thus reflects the difficulty level of the questions the student was able to answer correctly. Test scores range between 0 and 1,000. Upon completion of the test, students and their teachers get feedback on the achieved score.[1] The test results are important for students. After completion, students receive a certificate with their Stellwerk8 results. This certificate is usually provided to potential employers when students apply for an apprenticeship position in ninth grade.

In contrast to many other studies on peer effects in education (e.g. Angrist and Lang 2004, Carrell et al. 2016, Carrell and Hoekstra 2010, Hanushek et al. 2003, Kristoffersen et al. 2015, Lavy et al. 2012b), the data allow us to identify classroom composition, as we can identify the school, the track and the class teacher for each student (as in Burke and Sass 2013, Figlio 2007). We exclude data from special education institutions. We also exclude students with data entry mistakes, i.e. test scores which exceed the possible score range, implausible age and implausibly small or large classes.[2] In total, we drop 4.4% of test scores. Our final sample comprises 40,632 student-year observations with results for Math and German, 40,274 observations for English, and 39,818 observations for Science. Differences in observations arise from English and Science not being compulsory for all students. A year in our data is equivalent to a cohort as we only observe test scores in eighth grade. Test score distributions are approximately normal as shown in Figure A1 in the Appendix. On average, the students in our sample score 544 points in Math and 529 points in German. Besides test scores, the data contain information on gender, native

[1] Test scores are censored at 200 and 800 points before shown as feedback to students and teachers.

[2] Large classes are a sign that not the class teacher but an IT delegate entered the test scores into the archive. In this case, we cannot identify the true classroom composition of students. We have verified these cases by comparing the supposed teacher in our data against the professional register of teachers.

speaker status and date of birth. 49% of students are female, 14% speak a foreign language at home and the average age is 15 (see Table A1 in the Appendix). Classes comprise between 10 and 35 students, and between 0 and 15 SN students. Distributions of class aggregates are shown in Figure A2.

The data on class composition and test scores are linked with administrative records from the SPS. The SPS data contain students' diagnosis and counseling history, recorded by the caseworkers in the SPS database. Students enter the database if they had contact with the SPS before or during primary school. We observe the date of each contact per student, the reason for registration with the SPS, and information on diagnoses, treatment, therapy and other case-related data through caseworkers' comments. We classify students as special needs students if they got in contact with the SPS prior to secondary school. On average, 25.1% of all students and 4.3 students per class in eighth grade are classified as SN students. This is in line with the national average. The main reasons for registration with the SPS are learning difficulties, problems at school (e.g. teacher-student conflicts), or conspicuous behavior. On average, SN students score slightly worse than non-SN students in the standardized tests (cf. Figure A1).

Using administrative data from the SPS has several advantages. First, we observe directly whether a child has special needs, without relying on proxies for SN status such as gender, race, or parental background. Second, being an SN student is measured during primary school and independent of class composition in secondary education (within school-tracks). Third, the data allow measuring different degrees of special needs. Given the data register every contact with the SPS, it is reasonable to assume that children with more contacts have larger special needs than those with few contacts. For example, children with learning difficulties undertake several tests to assess their cognitive development, resulting in multiple contacts with the SPS. Children with conspicuous behavior have longer histories of contacts with potentially more complex therapy settings. We use the number of contacts for each child as an additional measure of the severity of special needs.

4 Empirical Strategy

We evaluate the impact of differential exposure to SN students on peer students' test performance. Although studying peer effects has a long tradition in economics, measuring such effects has proven challenging for three main reasons (Manski 1993). First, individuals in a group tend to behave similarly because they face similar environments. These correlated effects confound peer effects parameters, hindering the researcher from distinguishing true peer effects from common shocks shared by a group. The use of large data sets may resolve this problem by including a series of fixed-effects that control for unobserved heterogeneity at multiple levels (Burke and Sass 2013).

Second, it is difficult to separate the effect of the group on the individual from the effect of the individual on the group. This reflection problem exists because individual and peer outcomes are determined simultaneously. To resolve the reflection problem, applied work looks for proxies for peer quality that are then included in the regression equation. Examples of such proxies are gender and race (Hoxby 2000), student relocation (Angrist and Lang 2004), the presence of boys with feminine names (Figlio 2007), the presence of students retained previously (Lavy et al. 2012a), or the presence of children with family problems (Carrell and Hoekstra 2010, Kristoffersen et al. 2015). However, all these proxies only measure which peers are more likely to be low quality, without actually having precise information about peer quality. Using the detailed examination records of the SPS improves upon previous proxies for peer quality.

The third problem is that individuals tend to self-select into peer groups with specific characteristics. If this type of selection occurs, one cannot determine whether a change in outcome is a causal peer effect or a reason why individuals joined a group. In the context of peer effects in education, this self-selection issue can be resolved through quasi-randomization (Chetty et al. 2011, Lyle 2007, Sacerdote 2001) or by exploiting the natural variation in cohort composition over time (Carrell et al. 2016, Carrell and Hoekstra 2010, Hanushek et al. 2003, Hoxby 2000). The majority of papers exploiting the natural variation in cohort composition defines peers as the group of grade-level individuals at the same school. However, this definition is usually data-driven and not necessarily optimal. A natural approach is to define the peer group at the classroom level, because classroom peers exert a greater influence on individual outcomes (Burke and Sass 2013, Lazear 2001).

Our data allows us to identify classes and we combine both strategies—quasi randomization and natural cohort variation—to resolve the self-selection issue. In the canton of St. Gallen, the transfer from primary to secondary school is regulated, in the sense that students go to a given school depending on the district they live in. Schools may have different achievement tracks, and classes are strictly separated between tracks. Within each school-track, class formation is quasi-random, as students from different primary school districts are mixed and their SN status is not observed by secondary school administrators. Both primary schools and the SPS do not share information with secondary schools. To provide some supportive evidence in favor of the conditional randomization assumption, we follow the approach taken in Carrell and West (2010). For each class in a school-track, we resample 10,000 classes of the same size from the corresponding school-track and calculate the average exposure to SN students. For each of the sets of resampled classes, we calculate an empirical p-value as the proportion of simulations with exposure smaller than that observed in the original class. If class composition is random, the distribution of p-values within a school-track should be approximately uniform. We test against a uniform distribution using a one-sample Kolmogorov-Smirnov test. We reject uniformity one time out of 156. We take this as an indication that students' peer group at the school-track level is random.

Our identification relies on variation between classes within schools. Although families can potentially choose their district of residence and thereby influence schooling options for their children, possible selection into schools does not confound our results. Nevertheless, regions within the state of St. Gallen are very homogeneous in terms of demographics, indicating that such strategic behavior is most likely limited.[3] Mobility in Switzerland is generally low, about 80% of people do not move within five years (Eugster and Parchet 2013). Moving for school choice alone is likely to be a rare occurrence. However, even if it occurs, our identification strategy is not affected by mobility between school districts.

Formally, we are interested in the following linear model of student achievement:

$$TS_{icst} = \beta_0 + X'_{icst}\,\beta_x + \beta_{sn}\,SN_{icst} + \tilde{\beta}_{sn}\,\tilde{SN}_{icst} + \epsilon_{icst} \ , \tag{1}$$

where TS_{icst} represents the test score result of student i in class c at school-track s in year t. X is a vector of observable characteristics, which contains class size and binary indicators for gender, native speaker, age and whether a student receives additional tutoring outside the class environment. The binary variable SN_{icst} indicates whether student i is a child with special needs.

Our variable of interest is $\tilde{SN}_{icst}$, a measure of exposure to SN students in a given class. We calculate exposure for each student i as $\tilde{SN}_{icst} = \sum_{j\neq i} SN_{jcst}$ or $\tilde{SN}_{icst} = (n_{cst}-1)^{-1}\sum_{j\neq i} SN_{jcst}$, i.e., the number or proportion of SN students per class, excluding the SN status of i herself. In practice, excluding individuals' own status in their exposure measure has little influence on the empirical analysis. Our results are qualitatively and numerically very similar when using a simple count or share measure for all students per class. The peer spillovers parameter is $\tilde{\beta}_{sn}$, which represents the impact of SN students on their peers' school outcomes.[4]

Finally, ϵ_{icst} represents the error term, which we model in a components-of-variance framework. In detail, ϵ_{icst} is assumed to consist of two components: $\epsilon_{icst} = \mu_{st} + e_{icst}$, where μ_{st} is a school-by-track-by-year fixed effect and e_{icst} is an idiosyncratic error term.[5] This is equivalent to assuming $\mathrm{E}\left(e_{icst}|X, SN\right) = 0$, a common assumption in the literature (Cooley Fruehwirth 2009). This assumption would be violated if there is some residual input that is correlated with observable student characteristics. Given that class formation within

[3] The between municipality variation in the unemployment rate (coefficient of variation: 0.42), the share of rich (0.19) and poor taxpayers (0.69), the share with secondary (0.19), higher secondary (0.07) and tertiary education (0.22) is small.

[4] Our main specification does not include peers' characteristics aggregated at the class level ($\tilde{X}_{icst}$). This is because SN status is tied to some observables such as gender and age (boys and older children are more likely to be SN students.), and we are interested in the total effect of SN peers. We nevertheless perform a robustness check and show that the main results remain stable if we include class peer covariates.

[5] A less flexible, but widely used alternative specification is $\epsilon_{icst} = \phi_s + \gamma_t + r_{icst}$.

school-tracks is random and that we consider only predetermined student characteristics, the assumption appears plausible in our setting. The school-by-track-by-year component exploits the natural variation in cohort composition within school-tracks for each year in the most flexible way. We do not impose restrictions on μ_{st}. If the cross-cohort variation in students characteristics is random, this randomness implies that also $\mathrm{E}\left(\mu_{st} \mid X, SN\right) = 0$. Throughout the paper, we refer to equation 1 as the main model. In studying effect heterogeneity, we also consider more flexible versions of $\tilde{SN}_{icst}$ using cubic splines and also use an alternative measure of SN peer exposure based on an individual SN intensity proxy.

Being able to identify the class constitutes an advantage of our data, because, compared to the entire grade cohort, classmates are the more relevant peer group for the production of educational outcomes (Burke and Sass 2013, Carrell et al. 2009). This advantage also allows us to shed light on an additional challenge in peer effects analyses. Correlated effects can emerge due to possible non-random assignment of teachers to classes. To meet this challenge, we derive a model that controls for time-constant heterogeneity in achievement inputs at the teacher level. Revisiting the components-of-variance framework, we specify the error term as $\epsilon_{icst} = \nu_{cs} + \gamma_t + u_{icst}$, where ν_{cs} represents the teacher fixed effect. This latter specification filters the effects of class environment common to a teacher at a specific school and also potential matching between teachers and classes if it is time-constant. Any potential time trends affecting schools or teachers have to be captured by the common factor γ_t. In sum, while the former approach relies on variation between classes within a school-track each year, the latter relies on variation between classes of the same teacher in different years.

A trade-off exists between the identification within school-tracks and the identification within teachers. While the within-teacher identification has the potential advantage of filtering out correlated effects, this approach considerably reduces the variation that can be used for identification. Because teachers keep the same class throughout lower secondary school, they can teach an eighth grade class once every three years at most. We observe each eighth grade cohort between 2008 and 2016 and few teachers appear multiple times in the data. About 50% of teachers are observed exactly once, providing no identifying variation, and 29% are observed only twice. Therefore, to estimate the effect of SN peers, the within-teacher identification exploits much less information than the within school-track approach. It also assumes that there are no differential time/cohort effects for teachers or school-tracks. For these reasons, we keep specification 1 as the main model and defer the within-teacher specification to the sensitivity analysis.

Table 1: Estimates of SN class composition on student achievement

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	(a) All children (N = 40,632)							
	Math				German			
SN children (#)	−2.76***	−2.27***			−1.95***	−1.52***		
	(0.69)	(0.63)			(0.50)	(0.44)		
SN children (%)			−0.57***	−0.48***			−0.43***	−0.34***
			(0.14)	(0.12)			(0.09)	(0.08)
SN child	−14.56***	−11.79***	−14.84***	−12.05***	−12.99***	−7.54***	−13.26***	−7.76***
	(1.29)	(1.17)	(1.31)	(1.19)	(1.15)	(1.10)	(1.16)	(1.10)
	(b) Non-SN children (N = 30,408)							
	Math				German			
SN children (#)	−1.95***	−1.61***			−1.09**	−0.79*		
	(0.59)	(0.56)			(0.49)	(0.46)		
SN children (%)			−0.39***	−0.32***			−0.26***	−0.19**
			(0.11)	(0.10)			(0.09)	(0.08)
	(c) SN children (N = 10,224)							
	Math				German			
SN children (#)	−4.12***	−3.22***			−3.58***	−2.91***		
	(0.95)	(0.85)			(0.75)	(0.68)		
SN children (%)			−0.81***	−0.63***			−0.68***	−0.55***
			(0.17)	(0.15)			(0.13)	(0.11)
Individual covariates		✓		✓		✓		✓
Class size	✓	✓	✓	✓	✓	✓	✓	✓
School x track x year FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The table shows estimates for the effect of the number or share of SN children in a class on students' test scores in Math and German. Panel (a) considers all children, panels (b) and (c) condition on SN status. All models condition on school by track by year specific effects, identifying variation is provided by the variation of between classes within a school-track for each year. Class size is added as a control in all models, individual covariates are added as indicated. Standard errors clustered at the class level shown in parentheses. *, ** and *** denote $p < 0.1$, $p < 0.05$ and $p < 0.01$ respectively.

5 Results

5.1 Effects of classroom composition on test scores

Our main results are presented in Table 1. Panel (a) shows the effect of differential exposure to SN children for all other children in a class, controlling for individual SN status, panels (b) and (c) show the results for children without and with special needs. Columns (1)-(4) list the results for Math scores, columns (5)-(8) for German. All regressions control for class size and school-track-year fixed effects. We consider two measures of SN peers in a class, the number and the proportion of SN children.

We find that SN children score on average 14.5 points lower in Math and 13 points lower in German. This result is expected, as a large proportion of SN children have learning difficulties. The estimates also show that an additional SN child in a class reduces Math scores by about 2.7 points and German scores by about 2 points. Similarly, a one percentage point increase in SN children reduces test scores by between 0.6 and 0.4 points. All estimates are statistically significant at conventional levels. Comparing the results in panels (b) and (c), we find that the negative effect is mainly driven by the effects on children with special needs themselves, i.e. they tend to score lower if they are surrounded by other SN children in their class. Including student characteristics reduces the coefficients of interest slightly and increases their precision. The reduction in the coefficients is not statistically significant. Since the effects using the absolute and the relative measure of peer composition are similar, we focus on the count measure for the remainder of the paper for ease of interpretation.

The effects also persist for other subjects (see Table A2). Scores in English and Science are adversely affected by higher levels of SN children as well. We find that negative peer effects are strongest for Math and English. Our results also display the typical gender differences (not reported). Girls score significantly worse than boys in Math (37 points less) and Science (33 points less), but comparatively better in German and English (2 and 4 points more, respectively).

The results are also economically significant. Increasing the number of SN children in a class by one standard deviation (2.8 children) reduces Math test scores by 7.7 points, about 6.6% of the standard deviation in Math test scores. Scores of non-SN children would be reduced by 5.5 points, while those of SN children would be reduced by 11.5 points.

Our results remain qualitatively similar when controlling for average peer observables, i.e. adding $\tilde{X}_{icst}$ to the model, measuring class composition of characteristics such as gender and age (cf. Table A3). The estimated effects are slightly smaller. Since SN status is directly tied to some observables, i.e., SN children are less likely to be female and likely to be older, this observation is not surprising.

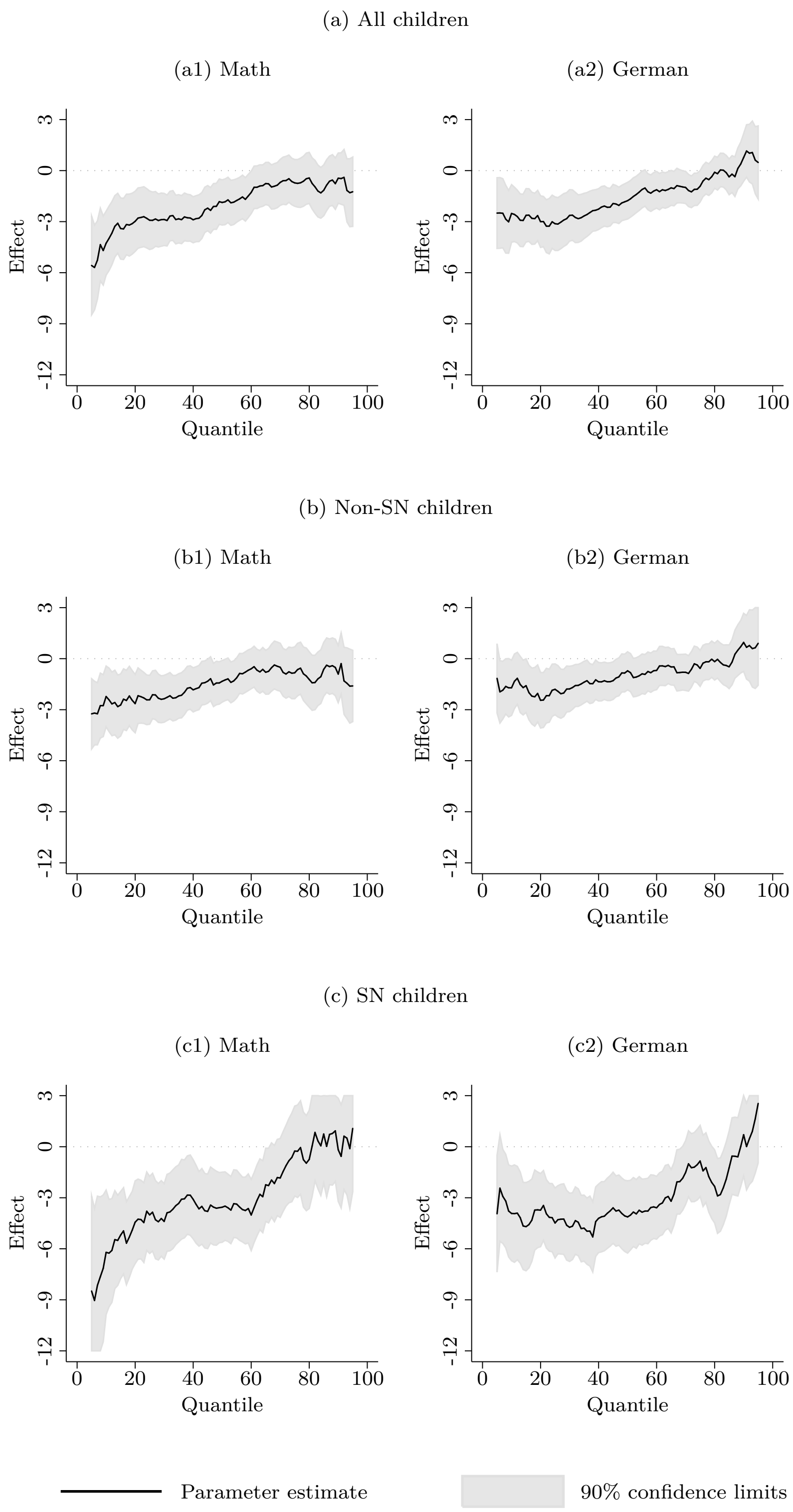
5.2 Effect heterogeneity

The previous results consistently show that test score of children with special needs themselves are impaired more by higher levels of mainstreaming than results of children without special needs. Effects might also be heterogeneous within these groups. In general, low achievement students may be more easily affected by disturbing influences in the classroom. To investigate this possibility, we transfer our main specification to a quantile regression framework and estimate the unconditional quantile treatment effects of SN peers over the distribution of test scores. We apply the unconditional quantile regression approach developed by Firpo et al. (2009), a standard approach to investigate heterogeneous effects (Borah and Basu 2013).

The results are shown in Figure 1, again separately for Math and German test scores and children with and without special needs. We consistently find that low achievement children are more strongly affected by SN children in their peer group compared to children at the upper end of the achievement distribution. This holds true for non-SN children and SN children. Within the group of non-SN children only those below the median are affected negatively. For SN children, the effect is larger and persists up to the 70%-quantile.

Having established that it is mostly the low achieving and more vulnerable children that suffer from higher levels of inclusion, we now focus on which children within the group of

Figure 1: Unconditional quantile regression, flexible, clustering at fe level, uniform scaling



Note: The figure shows estimates for the effect of the number of SN children in a class on students' test scores in Math and German for different quantiles of the test score distribution. Estimates are based on the unconditional quantile regression method of Firpo et al. (2009). Panel (a) considers all children, panels (b) and (c) condition on SN status. All models condition on school by track by year specific effects, identifying variation is provided by the variation of between classes within a school-track for each year. Class size is added as a control in all models, individual covariates are added as indicated. Standard errors are clustered at the school-track-year level. 90% confidence intervals are shown in grey.

special needs students are driving the effect. Our measure of special needs tags every child that has been in contact with the SPS during primary school or before, irrespective of the reason for the contact and the intensity of counseling. This measure might include very simple cases, such as requests for delaying school entry or cases where no further action was deemed necessary after diagnosing the child. Since we observe childrens' contact histories in the data, we construct a straightforward measure of individual *SN intensity* using the number of contacts with the SPS.

Table 2 shows estimates of peer effects exploiting the individual intensity measure. Using the sum of contacts as simple measure of overall classroom SN intensity (columns (1) and (3)), we again confirm that a higher degree of SN in a class is associated with lower student achievement. We then estimate the effects separately for the number of children in a class belonging to different quantiles of the individual SN intensity distribution. The results indicate that children with a low degree of SN have few, if any, statistically significant negative effects on peer achievement. However, having more classmates with pronounced SN, i.e., those at the upper end of the intensity distribution, reduces test scores significantly for all students. Again, the effect is more pronounced for those children who have some SN themselves. This result is consistent with the view that children with strong SN need more teacher attention at the cost of instructional time and can potentially disturb classroom learning to a significant extent (Lazear 2001).

The intensity measure also allows us to assess the sensitivity of our results to how we measure special needs. Using deciles of the individual SN intensity, we begin by considering all SN children in a cohort as having special needs and then gradually exclude from this definition the (next) lowest decile, until only children in the top decile of intensity are considered to have SN. Figure A3 shows the parameter estimates if we vary the definition of SN students in this way. Note that the definition of the children belonging to panel (b) or panel (c) is kept constant according to the baseline definition of SN children. The estimated effects are highly robust to the definition of SN children. As a general pattern, the estimates are slightly more negative and by construction become less precise if we increase the stringency of the definition of SN children. While our effect is robust to measurement choices, this exercise also shows a result similar to that in Table 2, i.e., that the negative effect is most likely driven by children with more severe problems during their childhood.

So far, our main results show that including more SN children in a class can be detrimental to student achievement. However, from our results we cannot make any statement about whether segregated education would be preferable. Presuming that inclusive education is a prior that cannot be changed, we try to address the question if there is some optimal composition such that negative spillover effects are mitigated. Returning to our initial definition of SN students, we look at effect heterogeneity along the dimension of SN students. We consider a more flexible version of our main model

Table 2: Estimates by SN intensity

	(1)	(2)	(3)	(4)
	(a) All children (N = 40,632)			
	Math		German	
SN intensity (Σ, class)	−0.22***		−0.19***	
	(0.06)		(0.04)	
Children within $[I_{min}, Q(0.2)]$ of intensity (#, class)		−1.50**		−0.83
		(0.73)		(0.58)
Children within $(Q(0.2), Q(0.4)]$ of intensity (#, class)		−1.05		−1.08
		(1.26)		(0.91)
Children within $(Q(0.4), Q(0.6)]$ of intensity (#, class)		−1.87*		0.24
		(1.00)		(0.76)
Children within $(Q(0.6), Q(0.8)]$ of intensity (#, class)		−2.90***		−2.16***
		(0.97)		(0.81)
Children within $(Q(0.8), I_{max}]$ of intensity (#, class)		−3.38**		−3.39***
		(1.38)		(0.95)
SN child	−9.88***	−9.62***	−6.09***	−6.08***
	(1.02)	(1.02)	(1.06)	(1.06)
	(b) Non-SN children (N = 30,408)			
	Math		German	
SN intensity (Σ, class)	−0.14**		−0.12**	
	(0.06)		(0.05)	
Children within $[I_{min}, Q(0.2)]$ of intensity (#, class)		−1.68**		−0.62
		(0.78)		(0.66)
Children within $(Q(0.2), Q(0.4)]$ of intensity (#, class)		−0.40		−0.67
		(1.32)		(1.07)
Children within $(Q(0.4), Q(0.6)]$ of intensity (#, class)		−1.26		1.52*
		(1.03)		(0.85)
Children within $(Q(0.6), Q(0.8)]$ of intensity (#, class)		−2.02**		−1.85**
		(1.03)		(0.91)
Children within $(Q(0.8), I_{max}]$ of intensity (#, class)		−2.01		−1.96*
		(1.30)		(1.03)
	(c) SN children (N = 10,224)			
	Math		German	
SN intensity (Σ, class)	−0.28***		−0.30***	
	(0.07)		(0.06)	
Children within $[I_{min}, Q(0.2)]$ of intensity (#, class)		−1.03		−1.66
		(1.08)		(1.01)
Children within $(Q(0.2), Q(0.4)]$ of intensity (#, class)		−2.13		−2.15*
		(1.66)		(1.26)
Children within $(Q(0.4), Q(0.6)]$ of intensity (#, class)		−2.54*		−1.83
		(1.44)		(1.17)
Children within $(Q(0.6), Q(0.8)]$ of intensity (#, class)		−4.60***		−2.80**
		(1.31)		(1.16)
Children within $(Q(0.8), I_{max}]$ of intensity (#, class)		−4.42***		−5.48***
		(1.70)		(1.27)
Individual covariates	✓	✓	✓	✓
Class size	✓	✓	✓	✓
School x track x year FE	✓	✓	✓	✓

Note: The table shows estimates for the effect of the overall sum of class intensity of SN and the effect of the number of children within quintiles of the individual SN intensity on students' test scores in Math and German. Panel (a) considers all children, panels (b) and (c) condition on SN status. All models condition on school by track by year specific effects, identifying variation is provided by the variation of between classes within a school-track for each year. Class size and individual characteristics are added as controls in all models. Standard errors clustered at the class level shown in parentheses. *, ** and *** denote $p < 0.1$, $p < 0.05$ and $p < 0.01$ respectively.

using cubic splines of $\tilde{SN}_{icst}$. Results are shown in Figure A4 for Math and Figure A5 for German. Again, we find that the incidence of the effect is largely on SN children. Overall, the linear approximation appears to be reasonable. However, looking at the marginal effect for the children driving the effect in panels (c2), we observe that negative spillovers worsen with the number of SN students and only start to impact student achievement after including more than 2–3 SN students in a class, or alternatively more than 15% of students in a class (results for shares not reported).

5.3 Changing the peer environment to the school cohort

One advantage of our data is that we can identify classroom composition, unlike many previous studies. Most of the interaction between students at school is expected to happen within the classroom, while interaction with other students in a cohort is limited to time before and after school or breaks.

To compare our results to the literature, we also perform the peer effects analysis on the school-grade level, ignoring classroom composition. Results are shown in Table A4 in the appendix. For comparability, the table presents results for shares of SN children. While the lower individual achievement of SN children status remains stable, the estimated peer effects are smaller compared to our main estimates. This shows that observing the relevant network structure is of key importance. Disruptive influences of classroom peers are likely to matter more for classroom production than influences from peers outside of the class environment.

5.4 The role of teachers

In the previously presented results, we used the identifying variation in the number of SN students per class across classrooms within school, track and years/cohorts. This is valid if two basic assumptions hold. First, students are randomly allocated to classrooms, i.e., without taking their SN status into account. Second, teachers are randomly allocated to classes. These assumptions are potentially violated if secondary school administrations receive information from primary schools on SN status of students. Classes with more difficult cases might then be assigned to more experienced or able teachers. This would most likely bias our results downwards, since experienced teachers improve achievement of all students (e.g. Rockoff 2004).

However, we are not aware of any formal information sharing arrangement between primary schools, secondary schools and the SPS when children transfer to secondary school and classes are composed. Without detailed knowledge of childrens' background, targeting teachers to classes is difficult. In addition, it is very unlikely within the institutional setting of Swiss schools that teachers would tolerate systematically being assigned to more troublesome and laborious classes.

Nevertheless, in order to investigate this possibility, we pursue the alternative identification strategy introduced in section 4. In our data, we can identify the teacher of each class. This allows us to estimate a model using only the identifying variation within a teacher over time, i.e., variation between classes in different years taught by the same teacher. Results are presented in columns (1)–(4) in Table A5 in the Appendix. We again find that SN children perform worse in the test. Non-SN children do not experience negative compositional effects once we control for the teacher. However, SN children themselves are still negatively affected by other SN children in the classroom. The estimated effects are slightly smaller than in our baseline results. This is consistent with Burke and Sass (2013), who also show that controlling for teacher impact reduces estimated peer effects. Since only about half of the teachers are observed teaching more than a single class within our sample period, and 30% teach only two classes, the identifying variation that can be used for identification is very limited, leading to imprecise estimates. If we repeat our main identification approach between classes within schools for the set of classes whose teachers are observed multiple times in the data (columns (5)–(8) of Table A5), we can replicate the within-teacher findings very closely. Nevertheless, we cannot fully dismiss the possibility that our main results are affected to some degree by residual selection, possibly based on teacher experience or intrinsic motivation as proxied by the willingness to serve as class teacher regularly.

6 Conclusion

We evaluate the effect of SN class composition on students' achievement in a standardized test for the student population of the Swiss region of St. Gallen between 2008 and 2016. To do so, we utilize a unique data set in which students' performance in a standardized test are matched to psychological examination records compiled by the School Psychological Service. Because SN status is determined in primary school, children have SN for a reason exogenous to their peers in secondary school. This feature allows us to estimate peer spillovers free from the reflection problem that has been difficult to overcome in the peer effects literature. With our data we can control for school-by-track-by-year specific effects and thus identify the spillovers by comparing classes with idiosyncratically high proportions of SN peers to classes in the same school-track cohort with idiosyncratically low proportions of SN peers.

We consistently find that higher levels of SN students in a class lower peers' achievement. The effect is persistent across all school subjects we observe. However, the incidence of negative spillovers is disproportionately on other SN students in a class and students at the bottom of the achievement distribution, while students who perform well are less affected.

This partly corroborates the result by Hanushek et al. (2002), who find that sharing a classroom with special education student does not necessarily impede achievement of regular education students. In addition, we demonstrate that the effect is mostly driven by individuals with severe SN, indicating that these students generate larger negative externalities for classroom education production. This holds true for all students regardless of SN status. There are important threshold effects of composition: Negative spillovers for other class peers are contingent on there being several SN students, indicating that disruptive influences exacerbate in conjunction. Our findings are robust to different specifications such as including peer characteristics aggregated at the class level, changing the peer environment to the school cohort, and estimating within-teacher effects.

This paper provides new insights on the bad apple principle in education (Carrell and Hoekstra 2010, Lazear 2001). In classroom education, the extent to which one student is able to learn during class time depends on the behavior of others in the class. If SN students take away learning time (e.g., through disruption or need for teacher's attention), one single SN student can impair the learning outcomes of all other students in the classroom. Our results suggest that the bad apple principle may emerge if sufficiently many SN students – 15% according to our estimates – are grouped in the same class.

The results have important implications for education policy. Whenever possible, schools and service providers should improve information exchange and early screening to deliberately guide class composition, such that SN students are distributed evenly across classes. If we take inclusion as given, an even distribution of SN students across classes should be preferred to policies in which some classrooms have anomalously large proportions of SN students. Teachers should devote particular attention to students in the lower half of the achievement distribution, as these are most likely to be affected by negative spillovers. Our results also indicate that classroom achievement would most likely improve if children with severe conspicuous behavior and a high propensity to disturb classroom learning were assigned to a separate educational environment. Other options could be special in-class counseling or part-time segregation. However, we cannot make a statement if such tracking is beneficial for the children in question or for the student population as a whole. Further research in this direction is needed. If SN tracking is also beneficial to special needs students, segregation should not be equalized to a denial of education opportunities.

References

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

Borah, B. J. and Basu, A. (2013). Highlighting differences between conditional and unconditional quantile regression approaches through an application to assess medication adherence. *Health Economics* 22(9), 1052–1070.

Burke, M. A. and Sass, T. R. (2013). Classroom Peer Effects and Student Achievement. *Journal of Labor Economics* 31(1), 51 – 82.

Carrell, S., Fullerton, R. L. and West, J. (2009). Does your cohort matter? Measuring peer effects in college achievement. *Journal of Labor Economics* 27(3), 439–464.

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

Carrell, S. E., Hoekstra, M. and Kuka, E. (2016). The long-run effects of disruptive peers. *Working Paper 22042*, National Bureau of Economic Research.

Carrell, S. E. and Hoekstra, M. L. (2010). Externalities in the classroom: How children exposed to domestic violence affect everyone's kids. *American Economic Journal: Applied Economics* 2(1), 211–28.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W. and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project Star. *The Quarterly Journal of Economics* 126(4), 1593–1660.

Cooley Fruehwirth, J. (2009). Can achievement peer effect estimates inform policy? A view from inside the black box. *Review of Economics and Statistics* 96(3), 514–523.

Eugster, B. and Parchet, R. (2013). Culture and Taxes: Towards Identifying Tax Competition. *Economics Working Paper Series 1339*, University of St. Gallen, School of Economics and Political Science.

European Agency for Development in Special Needs Education (2010). Inclusive education in action. *Report*, European Agency for Development in Special Needs Education.

Figlio, D. N. (2007). Boys named Sue: Disruptive children and their peers. *Education Finance and Policy* 2(4), 376–394.

Firpo, S., Fortin, N. M. and Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica* 77(3), 953–973.

Hanushek, E. A., Kain, J. F., Markman, J. M. and Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics* 18(5), 527–544.

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? *Review of Economics and Statistics* 84(4), 584–599.

Harrell, F. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis.* Springer.

Hoxby, C. (2000). Peer effects in the classroom: Learning from gender and race variation. *Working Paper 7867*, National Bureau of Economic Research.

Kristoffersen, J. H. G., Kraegpoeth, M. V., Nielsen, H. S. and Simonsen, M. (2015). Disruptive school peers and student outcomes. *Economics of Education Review* 45, 1 – 13.

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

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

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

Lyle, D. S. (2007). Estimating and interpreting peer and role model effects from randomly assigned social groups at West Point. *The Review of Economics and Statistics* 89(2), 289–299.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies* 60(3), 531–542.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review* 94(2), 247–252.

Sacerdote, B. (2001). Peer effects with random assignment: Results for dartmouth roommates. *The Quarterly Journal of Economics* 116(2), 681–704.

St. Galler Tagblatt (2016). Nur wenige Rekurse haben Erfolg. *St. Galler Tagblatt* .

U.S. Department of Education (2016). 38th annual report to congress on the implementation of the individuals with disabilities education act, 2016. *Report*, U.S. Department of Education, Office of Special Education and Rehabilitative Services.

Appendix: Tables and Figures

Table A1: Descriptive statistics

	Mean	SD	Median	Min	Max	N
SN child	0.25	0.43	0	0	1	40 632
SN children (#, per class)	4.34	2.87	4	0	15	40 632
SN children (%, per class)	25.11	17.74	21	0	100	40 632
Age at SPS registration	7.81	2.12	8	−2	16	8754
Female	0.49	0.50	0	0	1	40 632
Foreign	0.14	0.35	0	0	1	40 632
Tutoring	0.01	0.12	0	0	1	40 632
Age	14.93	0.69	15	13	17	40 632
Track: Real	0.35	0.48	0	0	1	40 632
Track: Sek	0.64	0.48	1	0	1	40 632
Track: UG	0.01	0.10	0	0	1	40 632
Year	2011.69	2.49	2012	2008	2016	40 632
Test score: Math	543.62	116.57	539	48	992	40 632
Test score: German	529.41	111.24	529	2	966	40 632
Test score: English	561.63	118.57	563	64	1023	40 274
Test score: Sciences	552.26	104.15	550	141	965	39 818
Class size	19.16	3.85	19	10	35	40 632

Table A2: Estimates of SN class composition on student achievement in different subjects

	(1)	(2)	(3)	(4)
	(a) All children			
	Math	German	English	Sciences
SN children (#)	−2.27***	−1.52***	−2.23***	−1.32**
	(0.63)	(0.44)	(0.51)	(0.53)
SN child	−11.79***	−7.54***	−16.10***	−3.16***
	(1.17)	(1.10)	(1.14)	(1.06)
	(b) Non-SN children			
	Math	German	English	Sciences
SN children (#)	−1.61***	−0.79*	−1.71***	−0.92*
	(0.56)	(0.46)	(0.49)	(0.54)
	(c) SN children			
	Math	German	English	Sciences
SN children (#)	−3.22***	−2.91***	−3.68***	−2.03***
	(0.85)	(0.68)	(0.76)	(0.72)
Individual covariates	✓	✓	✓	✓
Class size	✓	✓	✓	✓
School x track x year FE	✓	✓	✓	✓

Note: The table shows estimates for the effect of the number or share of SN children in a class on students' test scores in Math, German, English and Sciences. Sciences denotes the average score across Biology, Chemistry and Physics. Panel (a) considers all children, panels (b) and (c) condition on SN status. All models condition on school by track by year specific effects, identifying variation is provided by the variation of between classes within a school-track for each year. Class size and individual characteristics are added as controls in all models. Standard errors clustered at the class level shown in parentheses. *, ** and *** denote $p < 0.1$, $p < 0.05$ and $p < 0.01$ respectively.

Table A3: Estimates of SN students on student achievement controlling for average class characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	(a) All children (N = 40,632)							
	Math				German			
SN children (#)	−2.27***	−1.63***			−1.52***	−1.08***		
	(0.63)	(0.56)			(0.44)	(0.41)		
SN children (%)			−0.48***	−0.36***			−0.34***	−0.26***
			(0.12)	(0.11)			(0.08)	(0.08)
SN child	−11.79***	−11.30***	−12.05***	−11.55***	−7.54***	−7.24***	−7.76***	−7.48***
	(1.17)	(1.13)	(1.19)	(1.15)	(1.10)	(1.08)	(1.10)	(1.09)
	(b) Non-SN children (N = 30,408)							
	Math				German			
SN children (#)	−1.61***	−1.07**			−0.79*	−0.50		
	(0.56)	(0.52)			(0.46)	(0.44)		
SN children (%)			−0.32***	−0.22**			−0.19**	−0.14*
			(0.10)	(0.10)			(0.08)	(0.08)
	(c) SN children (N = 10,224)							
	Math				German			
SN children (#)	−3.22***	−2.65***			−2.91***	−2.34***		
	(0.85)	(0.79)			(0.68)	(0.68)		
SN children (%)			−0.63***	−0.54***			−0.55***	−0.45***
			(0.15)	(0.14)			(0.11)	(0.11)
Class peer covariates		✓		✓		✓		✓
Individual covariates	✓	✓	✓	✓	✓	✓	✓	✓
Class size	✓	✓	✓	✓	✓	✓	✓	✓
School x track x year FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The table shows estimates for the effect of the number or share of SN children in a class on students' test scores in Math and German. Panel (a) considers all children, panels (b) and (c) condition on SN status. All models condition on school by track by year specific effects, identifying variation is provided by the variation of between classes within a school-track for each year. Class size is added as a control in all models, individual covariates are added as indicated. All models also control for average class peer characteristics. Standard errors clustered at the class level shown in parentheses. *, ** and *** denote $p < 0.1$, $p < 0.05$ and $p < 0.01$ respectively.

Table A4: Changing the peer environment: Estimates of spillovers at the school cohort level

	(1)	(2)	(3)	(4)
	(a) ALL CHILDREN (N = 40,632)			
	Math		German	
SN children (%, within school)	−0.17**	−0.13*	−0.11	−0.10
	(0.08)	(0.08)	(0.07)	(0.07)
SN child	−10.99***	−10.97***	−7.11***	−7.11***
	(1.09)	(1.09)	(1.07)	(1.06)
	(b) NON-SN CHILDREN (N = 30,408)			
	Math		German	
SN children (%, within school)	−0.12	−0.09	−0.13*	−0.12
	(0.08)	(0.08)	(0.08)	(0.08)
	(c) SN CHILDREN (N = 10,224)			
	Math		German	
SN children (%, within school)	−0.36***	−0.32**	−0.14	−0.11
	(0.12)	(0.12)	(0.12)	(0.12)
School level covariates		✓		✓
Individual covariates	✓	✓	✓	✓
School size	✓	✓	✓	✓
School x track FE	✓	✓	✓	✓
Year FE	✓	✓	✓	✓

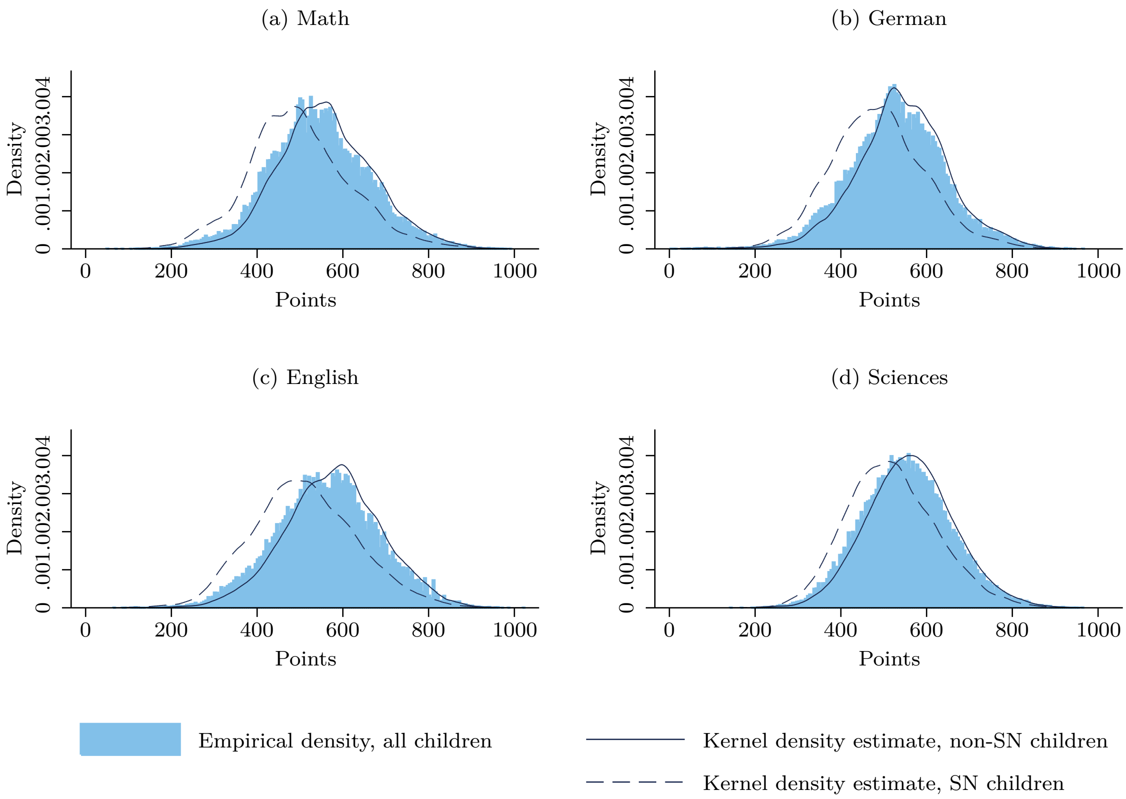
Note: The table shows estimates for the effect of the share of SN children in a school on students' test scores in Math and German. The treatment and peer group is defined at the school level to allow comparison to other studies. Panel (a) considers all children, panels (b) and (c) condition on SN status. All models condition on school by track and common year specific effects, identifying variation is provided by the variation within a school-track for different years. School size and individual characteristics are added as a control in all models, average school level covariates are added as indicated. Standard errors clustered at the school level shown in parentheses. *, ** and *** denote $p < 0.1$, $p < 0.05$ and $p < 0.01$ respectively.

Table A5: Alternative identification: The role of teachers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	(1) – (4) Within teacher identification				(5) – (8) Within school identification for teachers observed multiple times			
(a) All children								
	Math		German		Math		German	
SN children (#)	−0.13		−0.30		−0.24		−0.50	
	(0.38)		(0.39)		(0.54)		(0.45)	
SN children (%)		−0.03		−0.09		−0.11		−0.15*
		(0.07)		(0.07)		(0.11)		(0.08)
SN child	−9.36***	−9.38***	−6.20***	−6.28***	−9.78***	−10.00***	−7.37***	−7.57***
	(1.03)	(1.03)	(1.06)	(1.06)	(1.21)	(1.23)	(1.26)	(1.27)
(b) Non-SN children								
	Math		German		Math		German	
SN children (#)	0.27		0.25		0.00		−0.15	
	(0.40)		(0.43)		(0.54)		(0.49)	
SN children (%)		0.05		−0.01		−0.02		−0.06
		(0.07)		(0.08)		(0.10)		(0.09)
(c) SN children (N = 10,224)								
	Math		German		Math		German	
SN children (#)	−1.29**		−1.43**		−1.08		−1.28	
	(0.61)		(0.66)		(0.89)		(0.81)	
SN children (%)		−0.23**		−0.23**		−0.28*		−0.26**
		(0.10)		(0.11)		(0.16)		(0.14)
Individual covariates	✓	✓	✓	✓	✓	✓	✓	✓
Class size	✓	✓	✓	✓	✓	✓	✓	✓
School x track x teacher FE	✓	✓	✓	✓				
Year FE	✓	✓	✓	✓				
School x track x year FE					✓	✓	✓	✓

Note: The table shows estimates for the effect of the number or share of SN children in a class on students' test scores in Math and German. Panel (a) considers all children, panels (b) and (c) condition on SN status. All models condition on school by track by teacher specific effects, identifying variation is provided by the variation between classes taught in different years by the same teacher within a school-track. Class size and individual characteristics are added as controls in all models. Standard errors clustered at the class level shown in parentheses. *, ** and *** denote $p < 0.1$, $p < 0.05$ and $p < 0.01$ respectively.

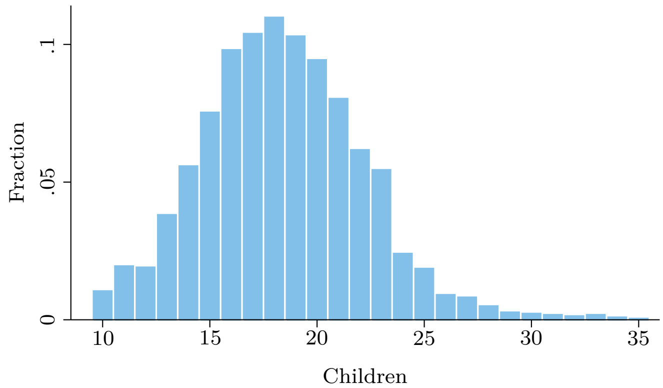
Figure A1: Test score distributions



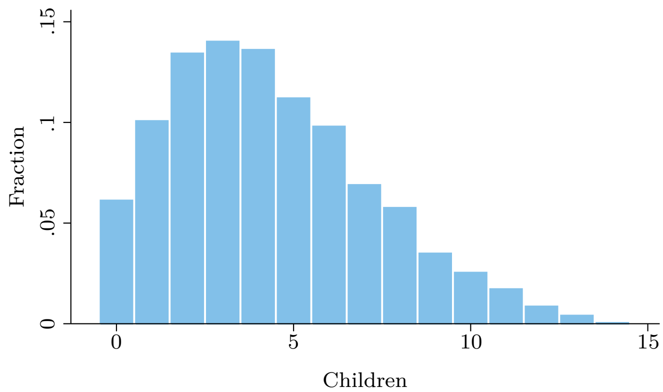
Note: The figure shows the distribution of test scores for Math, German, English and Sciences. The actual distribution of scores for all children is shown in blue in five point bins. The solid and dashed lines show kernel density estimates of the test score distribution for non-SN and SN children respectively, using an Epanechnikov kernel and a rule-of-thumb bandwidth.

Figure A2: Class distributions

(a) Class sizes

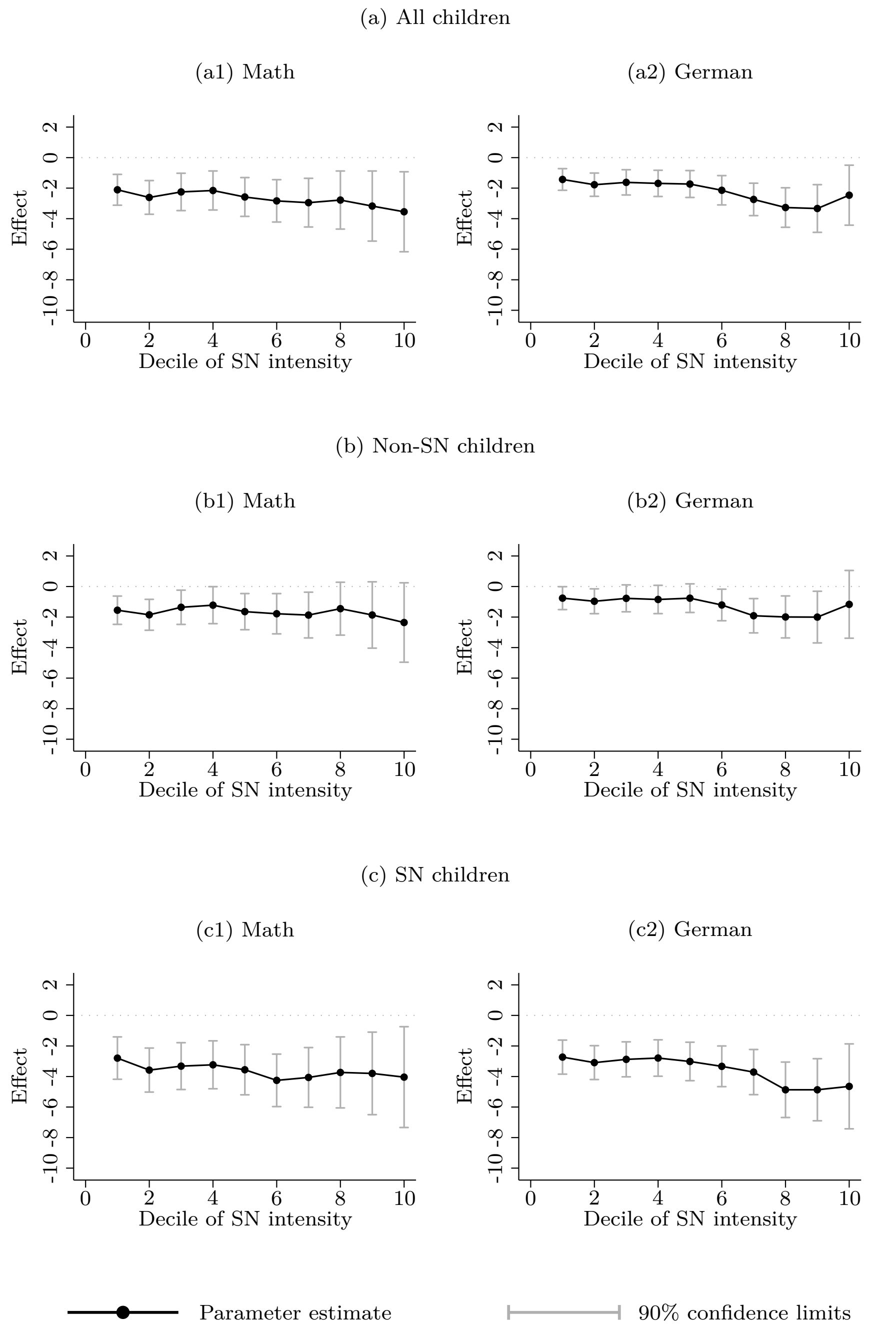


(b) SN children per class



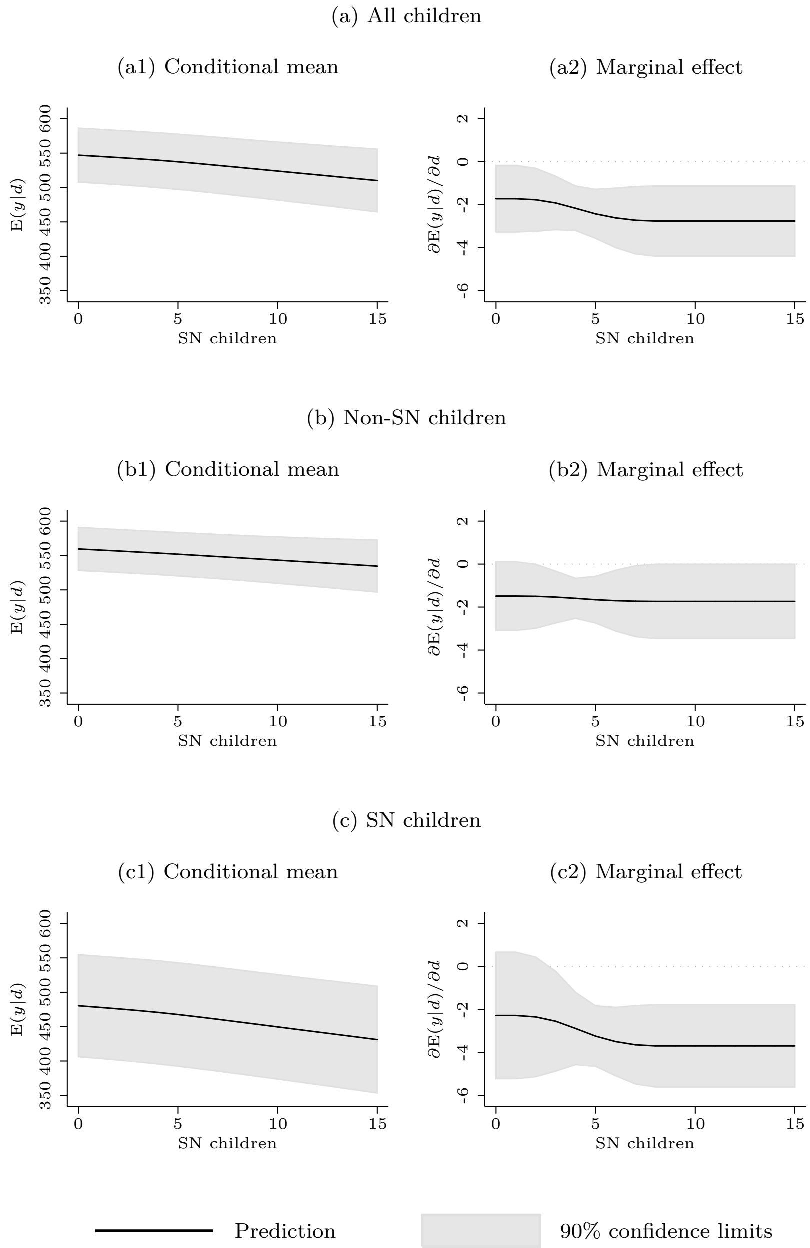
Note: The figure shows the distribution of class sizes and SN children per class for all classes in the main estimation sample.

Figure A3: Robustness: Measurement sensitivity of special needs definition



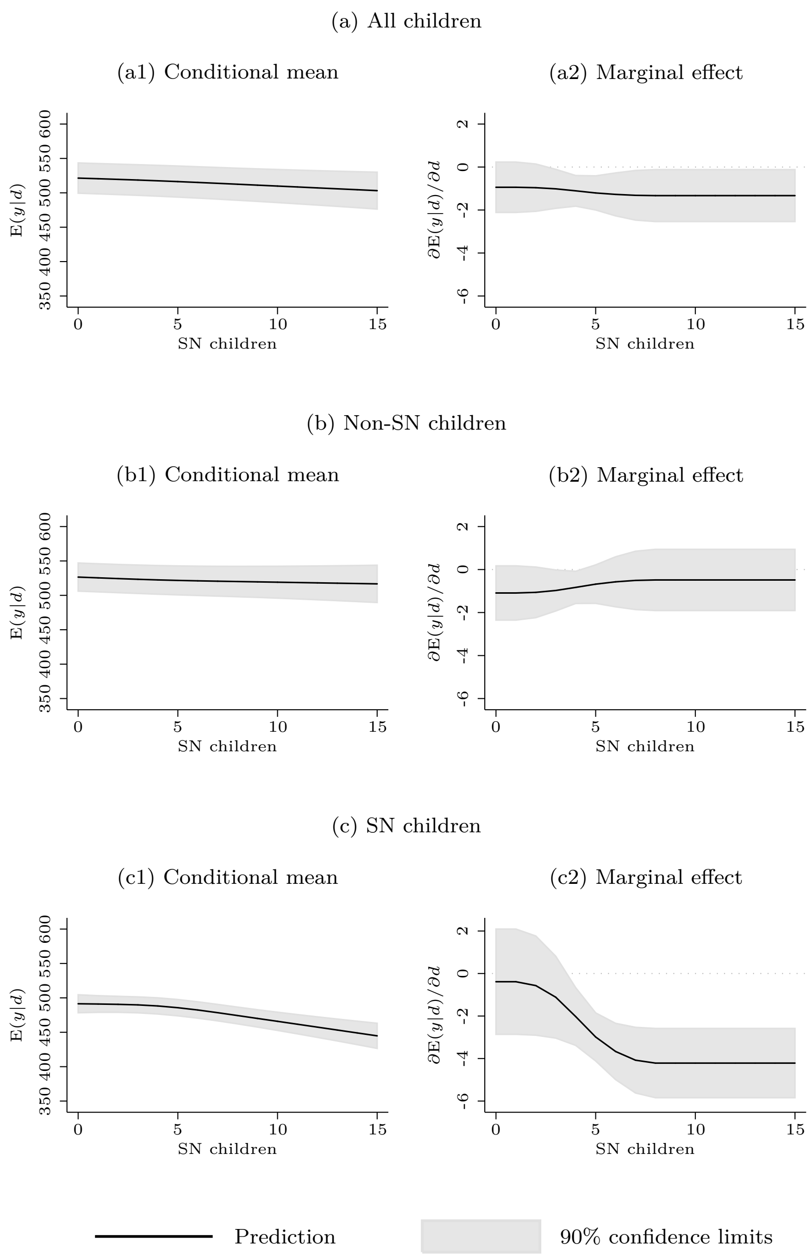
Note: The table shows estimates for the effect of the number or share of SN children in a class on students' test scores in Math and German. Estimates in each figure differ by the definition of SN children used. We vary the definition based on contact intensity. The leftmost estimate in each figure considers all children in a cohort ever in contact with the SPS as having special needs, regardless of the number of contacts. Moving to the right, the following estimates are based on gradually excluding from this definition the (next) lowest decile of the individual contact frequency distribution, until only children in the top decile of contacts are considered to have SN. Panel (a) considers all children, panels (b) and (c) condition on SN status. Note that the definition of the children belonging to panel (b) or panel (c) is kept constant according to the baseline definition of SN children. All models condition on school by track by year specific effects, identifying variation is provided by the variation of between classes within a school-track for each year. Class size is added as a control in all models, individual covariates are added as indicated. Standard errors are clustered at the class level. 90% confidence intervals are shown in grey.

Figure A4: Flexible estimates of SN class composition on student achievement in Math



Note: The figure shows flexible estimates of the conditional mean function and the marginal effect of the number of SN children in a class on students' test scores in Math. Estimates are based on regressions including restricted cubic splines of the number of SN children per class. Knots are set at percentiles 10, 50 and 90 following Harrell (2015). Panel (a) considers all children, panels (b) and (c) condition on SN status. All models condition on school by track by year specific effects, identifying variation is provided by the variation of between classes within a school-track for each year. Class size is added as a control in all models, individual covariates are added as indicated. Standard errors are clustered at the school-track-year level. 90% confidence intervals are shown in grey.

Figure A5: Flexible estimates of SN class composition on student achievement in German



Note: The figure shows flexible estimates of the conditional mean function and the marginal effect of the number of SN children in a class on students' test scores in German. Estimates are based on regressions including restricted cubic splines of the number of SN children per class. Knots are set at percentiles 10, 50 and 90 following Harrell (2015). Panel (a) considers all children, panels (b) and (c) condition on SN status. All models condition on school by track by year specific effects, identifying variation is provided by the variation of between classes within a school-track for each year. Class size is added as a control in all models, individual covariates are added as indicated. Standard errors are clustered at the school-track-year level. 90% confidence intervals are shown in grey.