Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1474

Does in-service training in special education increase the effectiveness of general classroom teachers? Evidence from the large-scale "Special Education Pedagogy for Learning" program in Sweden

Leah Natasha Glassow
Goteborgs Universitet

Nils Kirsten
Uppsala University

This study evaluates the impact of a large-scale professional development program for general classroom teachers in Sweden, the "Special Education Pedagogy for Learning" (SFL) program. Linking program administrative data to national full population register data, we apply a dynamic difference-in-differences estimation of school-level program participation on grade 6 and grade 9 student mathematics and Swedish grades and national test scores. The findings indicate that the program did not lead to improved achievement in national test scores when effects were averaged across all students. However, heterogeneity analyses suggest moderate effect sizes (0.077 standard deviations) in Swedish national test scores for newly arrived childhood immigrants. Conversely, we do not find any effect of the program for students with individualized education programs across any outcome of interest.

VERSION: May 2026

Suggested citation: Glassow, Leah Natasha, and Nils Kirsten. (2026). Does in-service training in special education increase the effectiveness of general classroom teachers? Evidence from the large-scale "Special Education Pedagogy for Learning" program in Sweden. (EdWorkingPaper: 26-1474). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/p0s9-9131

Does in-service training in special education increase the effectiveness of general classroom teachers? Evidence from the large-scale "Special Education Pedagogy for Learning" program in Sweden

Abstract

This study evaluates the impact of a large-scale professional development program for general classroom teachers in Sweden, the "Special Education Pedagogy for Learning" (SFL) program. Linking program administrative data to national full population register data, we apply a dynamic difference-in-differences estimation of school-level program participation on grade 6 and grade 9 student mathematics and Swedish grades and national test scores. The findings indicate that the program did not lead to improved achievement in national test scores when effects were averaged across all students. However, heterogeneity analyses suggest moderate effect sizes (0.077 standard deviations) in Swedish national test scores for newly arrived childhood immigrants. Conversely, we do not find any effect of the program for students with individualized education programs across any outcome of interest.

Keywords: large-scale teacher professional development, special and inclusive education, special educational needs, childhood immigrants, event study, difference-in-differences

1 Introduction

High quality teaching has long-lasting impacts on student outcomes (Chetty et al., 2014)—and consequently, in Sweden and many countries around the world, government bodies have employed large-scale teacher professional development (PD) programs in response to declining test scores (Dee & Jacob, 2011; Grönqvist, Öckert, & Rosenqvist, 2025; Holmlund, Häggblom, & Lindahl, 2024). Many studies have evaluated programs targeting vulnerable student groups, including those with a foreign language or immigrant background, as well as low-achieving students more broadly (Gamse et al., 2008; Jacob & Lefgren, 2004; Machin & McNally, 2008; Yoshikawa et al., 2015). Less often, however, have large-scale PD programs targeting students with special educational needs or using special or inclusive educational approaches been rigorously evaluated (Donath et al., 2023).

On-the-job training in this endeavor may provide much needed support to teachers, given that many students with special educational needs attend general classrooms (Cooc, 2019; OECD, 2015a; OECD, 2021). Additionally, expanding migration flows have led to increasingly linguistically and culturally diverse classrooms in many countries, including in Sweden (Pinson, Bunar & Devine, 2023).

Although the evidence for supplementary teaching and special education designation is strong (Coffey et al., 2026; Gersten et al., 2020; O'Hagan & Stiefel, 2024; Schwartz et al., 2021), the ways in which general education teachers can be aided to support students with special educational needs or multilingual backgrounds are less clear (Bal & Perzigian, 2013; Carroll et al., 2017; Fuchs et al., 2015; Fuchs et al., 2025). Some special education researchers have advocated that general education teachers choose methods that avoid placing unnecessary demands on skills like working memory and phonological processing, which some students may struggle with (Fuchs et al., 2016). Others advocate for within-class differentiation in which students with varying needs may be directed to learn

different content, using different methods, working in different learning environments, or demonstrating their knowledge in different ways (Kahmann et al., 2022). In contrast, still others caution that excessive within-class differentiation may undermine inclusion and instead advocate common instructional goals, combined with responsive teaching and integrated forms of support (Florian, 2019; Jordan, Schwartz, & McGhie-Richmond, 2009; Kalinowski, Gronostaj, & Vock, 2019).

Teacher survey responses as well as qualitative evidence from a range of different countries show that teachers in general classrooms struggle to implement special and inclusive education pedagogical approaches and struggle to support students with special educational needs or multilingual backgrounds in their classrooms (Bölte et al., 2021; Cooc, 2019; Deunk et al., 2018; Eurydice, 2019; Saivoleinen et al., 2012; Söderlund et al., 2024). Moreover, though many PD programs have aimed to support students with special educational needs, such as reading difficulties, dyslexia or ADHD (e.g., Chao et al., 2017; Didion et al., 2021; Scanlon et al., 2008; Zentall & Javorsky, 2007), causal evidence of more general special or inclusive education PD is rare—and all the rarer for programs implemented at scale (Donath et al., 2023).

In this study, we address this gap by evaluating the effectiveness of one such inclusive education PD program: the large-scale "Special Education Pedagogy for Learning" (hereafter, SFL) program, which was launched in 2016 and formally ended in 2025. SFL represented a national policy effort to embed special and inclusive education principles within general classroom teaching and thereby strengthen teacher capacity to address diverse student needs across a range of subjects. Tasked by the Swedish government, both the National Agency for Education (SNAE) and the National Agency for Special Needs Education (SPSM) were responsible for the development and rollout of the SFL program, in which about 60,000 Swedish teachers participated, and which cost 680 million SEK (roughly 70 million USD, SNAE, 2025).

Via a dynamic difference-in-differences estimation strategy (Callaway & Sant'Anna, 2021) and by linking program administrative data to national full population register data across 4936 schools and just under 2 million students, we exploit the staggered rollout and determine whether the positive effects of SFL, should they exist, persisted over time or tapered off, as our data extend up to 8 years prior to and up to 6 years after the SFL program intervention. The main outcomes of interest are grades and national test scores in mathematics and Swedish reported in the national register at Statistics Sweden. We test average effects of SFL as well as program effects for students with higher likelihoods of benefitting from the intervention, such as those with individualized education programs, a foreign language background, or those newly arrived in Sweden. We also test heterogeneity based on gender, school achievement quartiles, average teacher qualification levels in the school, and school type. Finally, we conduct a series of sensitivity analyses to explore the robustness of the findings, using alternative model and treatment specifications, and examining grading leniency, program dosage, student compositional changes, and cohort-specific effects.

The findings indicate that the SFL program produced statistically significant but very small effects on Swedish grades, but no effects in mathematics grades or in either subject in the national test scores, implying unstable effects or that the SFL program may have increased grading leniency.

Heterogeneity analyses suggested no effect of the program for students with individualized education programs across any outcome of interest. However, for childhood immigrants arriving in Sweden within 5 years, we find improvements in Swedish national test scores of 0.077 standard deviations but no changes in school grades. We also find higher test scores for students in low-achieving schools in Swedish (0.05 standard deviations), though these do not reach conventional levels of statistical significance. For school-level characteristics, we find that participating schools with low shares of certified teachers and

independent schools showed increased mathematics national test scores (between 0.06 and 0.075 standard deviations). Finally, concerning dynamic effects, the event study results indicate that even in the cases where subgroup analyses indicated positive effects, these were unstable in the post-treatment years. While the difference-in-differences design seeks to recover causal effects by comparing outcome trends between participating and non-participating schools, causal certainty is weaker than under randomized participation. Nevertheless, the estimated effects are robust across multiple specifications.

Beyond assessing the effectiveness of the large public investment in the SFL program, our study advances several strands of literature which we outline from most to least proximal. First, despite the proliferation of PD programs aimed at supporting students with special educational needs, rigorous evaluations are rare and typically small in scale. For example, while a meta-review by Donath et al. (2023) included 324 studies, only 17 percent of these had an experimental or quasi-experimental design and the review did not differentiate results from these studies from the cross-sectional or single-group studies. Similarly, a recent review of experimental and quasi-experimental studies on special education PD by Didion et al. (2026) identified only 9 studies which included effects on students. Evidence is further fragmented with reviews of differentiated-instruction PD documenting highly variable effects across programs and little explanation for why (Kahmann et al., 2022). Our study thus provides a rare evaluation of a large-scale special/inclusive education PD program targeted towards general classroom teachers, adding comparatively more robust evidence to a literature in which effects have seldom been assessed through high-quality research designs.

Second, we contribute to the empirical literature on the effectiveness of inclusive education practices within general classrooms. Across reviews, research on how best to support diverse learners in mainstream settings remains inconclusive. Research syntheses of differentiated instruction show that, although some studies report small to moderate positive

effects, the evidence base is heterogeneous and difficult to generalize because implementations, outcome measures, and contrasts vary widely (Deunk et al., 2018; Smale-Jacobse et al., 2019). Similarly, Carroll et al. (2017) note that while strong evidence exists for targeted interventions, much less is clear regarding effective everyday teaching and adaptations. In contrast, credible evidence points to the effectiveness of intensive, supplemental, and highly explicit instruction delivered in small groups or one-to-one formats (Fuchs et al., 2025). Our findings align with this broader pattern: despite students with individualized education programs being a stated target group of the SFL program, we observe consistently null or slightly negative effects for these students across all outcomes, persistent across a range of sensitivity checks.

At the same time, the results indicate that the program improved Swedish national test scores for newly arrived childhood immigrants, another group emphasized in the program's aims. Although a wide range of policies and practices are promoted across OECD countries to support immigrant students' integration (Choi, Mao & Park, 2025; Choi & Lee, 2020; Eurydice, 2019; OECD, 2015b), the existing knowledge base is dominated by descriptive or correlational studies and provides little rigorous causal evidence on which pedagogical approaches actually improve academic outcomes for this group (Bal & Perzigian, 2013). Our findings suggest that the SFL program may have benefited some student groups—such as newly arrived immigrants—while offering little measurable academic support for students with individualized education programs.

Third, the mostly null effects from our study add to a growing body of research indicating that maintaining robust effects of PD programs at scale is challenging. Reviews such as Kraft et al. (2018) interpret the weaker results of large-scale initiatives as stemming from lower intensity, reduced targeting, and substantial implementation heterogeneity. The dynamic estimates in our study show that even the positive subgroup effects found (e.g., for newly arrived immigrants) tapered off within the post-treatment period. Together, these findings

reinforce the difficulty of achieving and maintaining robust instructional improvements through large-scale PD models. They also complement evidence from prior national PD programs in Sweden, which similarly displayed small effects (0.01 to 0.03 SD), albeit reaching statistical significance (Grönqvist et al., 2025; Holmlund et al., 2024).

The article is structured as follows. In section 2 we present an overview of the context and content of the SFL program in Sweden, section 3 presents the data sources, treatment definition(s), and balancing and pre-trend tests, as well as the outcome variables. Section 4 presents the main identification strategy, heterogeneity analyses, and planned sensitivity checks. In section 5, we present the main results for the pooled analyses, heterogeneity analyses, and discuss the results from a battery of sensitivity checks which are documented in detail in the supplementary materials. Section 6 discusses and contextualizes the findings and provides some guidelines for future PD oriented policy efforts.

2 The SFL program

2.1 Context

Between the years 2003 and 2012, student performance in Sweden declined more rapidly than in any other country participating in PISA (OECD, 2015a). Moreover, in 2014, almost 15% of students in Sweden failed to qualify for upper-secondary school beginning at the age of 16—and overrepresented among this group were students with special educational needs, students with foreign language backgrounds, and newly arrived students (Regeringen, Regeringsbeslut U2015/05783/S). In response to this decline in achievement, the Swedish government tasked the SNAE with developing a series of large-scale professional development programs, including the Boost for Mathematics (BFM), the Boost for Reading (BFR), and the SFL program which began in 2016.

Like both BFM and BFR, the SFL program was large in scale—and the Swedish government allocated approximately 680 million SEK (roughly 70 million USD) towards its completion. Moreover, if each participating teacher spent at least 60 hours on SFL activities and roughly 60,000 teachers took part between 2016 and 2025, the program required about 3.6 million teacher-hours during this period. Participation in the SFL program was voluntary and occurred during regular school working hours, supported through a targeted state grant administered by the SNAE. The funding was allocated directly to participating schools to compensate for the time that coaches spent leading the collegial learning cycles. Depending on whether the school applied for participation with one or two teacher groups, the grant covered either 10 percent or 20 percent of the coach's employment position over one academic year.

The SNAE states in their description of the program that it does not advocate any one particular teaching method. Rather, the aim is to provide special education-oriented material that aids professional dialogues in schools about how to improve their work. The content of the SFL program reflects an ambition to teach students with a broad range of educational needs in the general classroom, including students with formal special educational needs, language difficulties, or who have recently immigrated to Sweden (SNAE, 2025). In fact, the program explicitly states that all students form the target group, though the formal budgetary justification for SFL lists students with special educational needs and students with a foreign language background as target groups.

One large (Thoutenhoofd et al., 2020) and three small (SNAE, 2025) evaluations of SFL have been conducted, all of them focusing on teacher attitudes and none of which involved a control group. The largest evaluation (Thoutenhoofd et al. 2020) used pre- and post-questionnaires from over 3,000 teachers, along with observations and interviews in 30 schools, to examine changes in attitudes, self-efficacy, and collaborative practices related to inclusion, as well as to analyze program content. Teachers generally reported more positive views of

inclusive education and valued the collegial discussions, though many found the program content difficult to translate into practice. The three smaller evaluations also reported positive views on the program among supervisors and participants (SNAE, 2025).

The SFL program has also been criticized in public debate especially for its mention of Universal Design for Learning (UDL), allegedly for lacking a firm evidence base (Expressen, 2023). Nevertheless, the SFL program was broad in scope, and included modules ranging from inclusion and participation, multilingualism, neurodevelopmental disorders, attention and the learning environment, and mathematics didactics, among others, which we expand upon in the following section.

2.2 SFL program content

In line with past nationally implemented PD programs in Sweden, SFL adopted the so-called 'collegial learning' model developed by SNAE in 2011 (Kirsten & Carlbaum, 2020). This model borrows elements from several professional development frameworks, including lesson study (Chen & Zhang, 2019), learning study (Pang & Ling, 2012), and Nordic professional development models such as 'research circles' (Holmstrand & Härnsten, 2003). In this collaborative, practice-oriented approach, the aim is that teachers improve their teaching practices in groups taking part in learning modules, via individual study, joint lesson planning, discussion, and time to test these practices within their classrooms and reflect on their level of success. For an overview of the 'collegial learning' approach where modules were divided into used in SFL, see Figure 1.

Each participating teacher group took part in SFL for one academic year. During that year, modules were undertaken sequentially, with compulsory schools needing to complete at least 2 modules per year. If schools did not implement the required number of modules, they could be forced to pay back the state administered grant. At participating schools, one or more

teacher groups (on average consisting of 10 teachers, see Table 2) met weekly throughout the academic year. The sessions followed structured modules provided on the program website. Program materials included theoretical readings, videos, examples, and step-by-step guidance for both individual preparation and teacher meetings. Each module followed a structured cycle of (a) reading, (b) joint lesson planning/collegial discussion, (c) classroom application, and (d) joint reflection (see Figure 1). Schools selected which modules to engage with based on local needs, though all needed to complete at least 1 of 3 of the foundational modules on inclusion. Typically, the school principal selected participating teachers, and the principal, the coach, and/or the teacher group decided which modules to complete.

A central component of SFL was the use of trained coaches who guided the teachers' professional learning. Although the module content was housed within an online learning portal, the coaches and weekly meetings ensured a high level of face-to-face teacher training. Coaches were usually special education teachers, but certified teachers without a special education degree were also eligible provided they received additional training from the SNAE, though all coaches were offered training (SNAE, 2025). The coaches' role was to lead teacher discussions according to instructions specified in the module content. While the coaches were typically employed at the school, this was not always the case, and sometimes groups of several schools collaborated together via the principal organizer (a municipality or the owner of the independent school[1]).

The materials of the SFL program were developed based on cooperation between the SNAE, the SPSM, and commissioned university researchers. The program materials included a wide variety of inclusive and special educational approaches and topics, though the vast majority had a strong focus on inclusive education topics and strategies. For descriptions

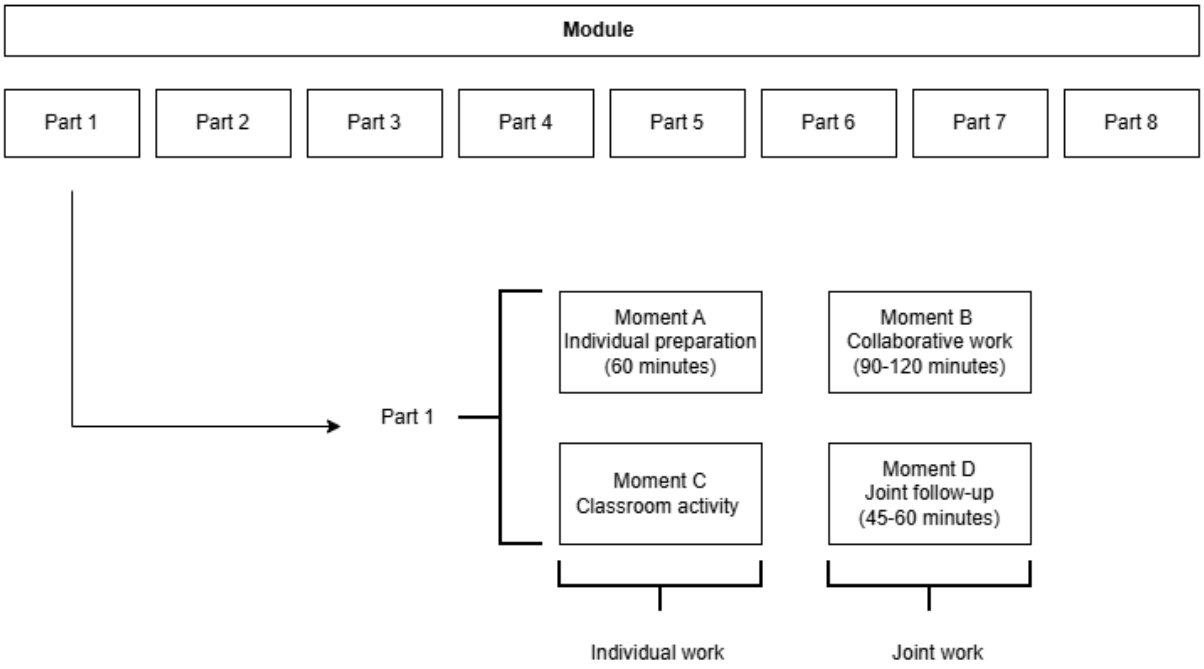
[1] In Sweden, independent schools may not charge fees for their students but can apply their own admission rules based on residential proximity or capacity.

of select modules, see Table 1. Examples of some of the intended outcomes of the foundational and mandatory modules on inclusion and participation include increased ability to conduct accessible and adapted teaching, strengthened student-teacher relationships, and a safer and more friendly school climate, which in turn, according to the program, should foster learning (SNAE, 2025).

Examples of optional modules included 1) Participation in the Learning Environment, 2) Inclusion in School Practice, 3) Attention, Interaction, and Communication, 4) Multilingualism, 5) Neurodevelopmental Conditions, 6) Mathematics Didactics and Special Education, 7) Accessible Learning with Digital Tools, and 8) Systematic Work with Support in Teaching. Some explicit aims of these modules state for example, increased awareness of: how the environment shapes conditions for participation and strategies to foster a supportive learning environment, how varied instruction can meet diverse student needs, methods and strategies which support learning for children with attention difficulties, methods to promote motivation and school attendance, and methods to support students with language difficulties either due to a foreign language background or delays in second-language acquisition (SNAE, 2025).

Figure 1

The SFL program's 'collegial learning' model: module structure



Note. The figure shows the structure of SFL within an example module, which is divided into 8 subparts in which the cycle between moments (a) through (d) occur.

Table 1

Overview of content in select SFL modules provided by the SNAE

Module	SNAE description
Inclusion and participation (foundational)	The module aims to provide in-depth knowledge and understanding of what inclusion entails and how an inclusive learning environment and students' participation influence learning. The module offers tools for reflecting on, observing, testing, analysing, and discussing one's own teaching practice and organisational context. Furthermore, the module aims to develop the ability to design accessible and adapted instruction.
Interaction for inclusion (foundational)	A starting point of the module is the idea that strong relationships and effective communication form the foundation of a school that promotes well-being and learning for all students, as well as for school staff. Teachers' continuous learning and collaboration to improve and develop the school (collective learning) are foregrounded. Special educational challenges are viewed as a collective responsibility rather than as individual problems. When teachers jointly develop their communication and relational competence across the whole school, key teacher-student relationships can be strengthened, a safe and supportive social climate can be fostered, and a positive learning environment can be established.
Participation in the learning environment	The module is based on a model of participation that supports identifying, reflecting on, discussing, analysing, and developing the learning environment. Well-developed work on participation can support the creation of a successful learning environment that includes all students. Increased awareness and knowledge of how the learning environment shapes conditions and sets boundaries for students' opportunities for participation also enhances the potential for change and development.
Attention, interaction and communication	This module aims to provide in-depth knowledge of how teaching and learning environments can be designed in relation to students' variation in perception, cognitive abilities, executive functions, and communication. The content enhances understanding of the pedagogical consequences that difficulties in different areas may entail, while also presenting approaches, methods, and attitudes that support learning and interaction. The module also aims to develop and deepen professional relational competence that promotes motivation and school attendance.
Multilingualism	The module aims to provide in-depth knowledge and understanding of how teaching and learning environments can be adapted to students' variation from a multilingual perspective. The content presents approaches, methods, and attitudes that support the learning conditions of multilingual students, both in relation to expected second-language development and in cases where second-language development is delayed.
Neurodevelopmental conditions and accessible education	The purpose of the module is to enable teachers and staff involved in teaching to develop their knowledge of neurodevelopmental conditions. The module presents selected research within the neuropsychiatric field that provides guidance on how teaching and learning environments can be designed in relation to students' needs and prerequisites.

Note. The figure displays select modules from the SFL program with modules summaries provided by the SNAE (SNAE, 2025).

3 Data and empirical approach

3.1 Data sources

The study combines administrative data from the Swedish National Agency for Education on SFL program participation between the years 2016 and 2022 and full population register data from several national population registers from Statistics Sweden between 2014 and 2022. The first is the pedagogical register which contains individual level annually updated information on all individuals employed in Swedish schools. It includes teacher credentials, age, gender, employment information, position and role. We also link information from the pupil register, the grade point register and the national test register, which includes information on student migration background, grades and national test scores, as well as whether students have individualized education programs. The teacher and student data are linked at the school level, meaning although we cannot link individual students to teachers, we can identify school-level treatment effects. The school-level variables include its unique identifier, location, type (independent or public), and size.

We use data for each participating and never-participating compulsory school across 9 academic school years (2014/2015-2022/2023), amounting to 1,967,114 individual students in 4936 schools, as well as the full population of teachers (whose characteristics are averaged at the school level) in these schools.

As is typical in large-scale administrative datasets, some variables contain missing observations. Missingness is negligible for core demographic variables (e.g., gender and individualized education programs) and modest for background and outcome measures such as parental education and national test scores (approximately 5-10 percent). Because the unit of treatment is the school and missingness occurs at the individual level, this is unlikely to substantively affect school-level aggregates.

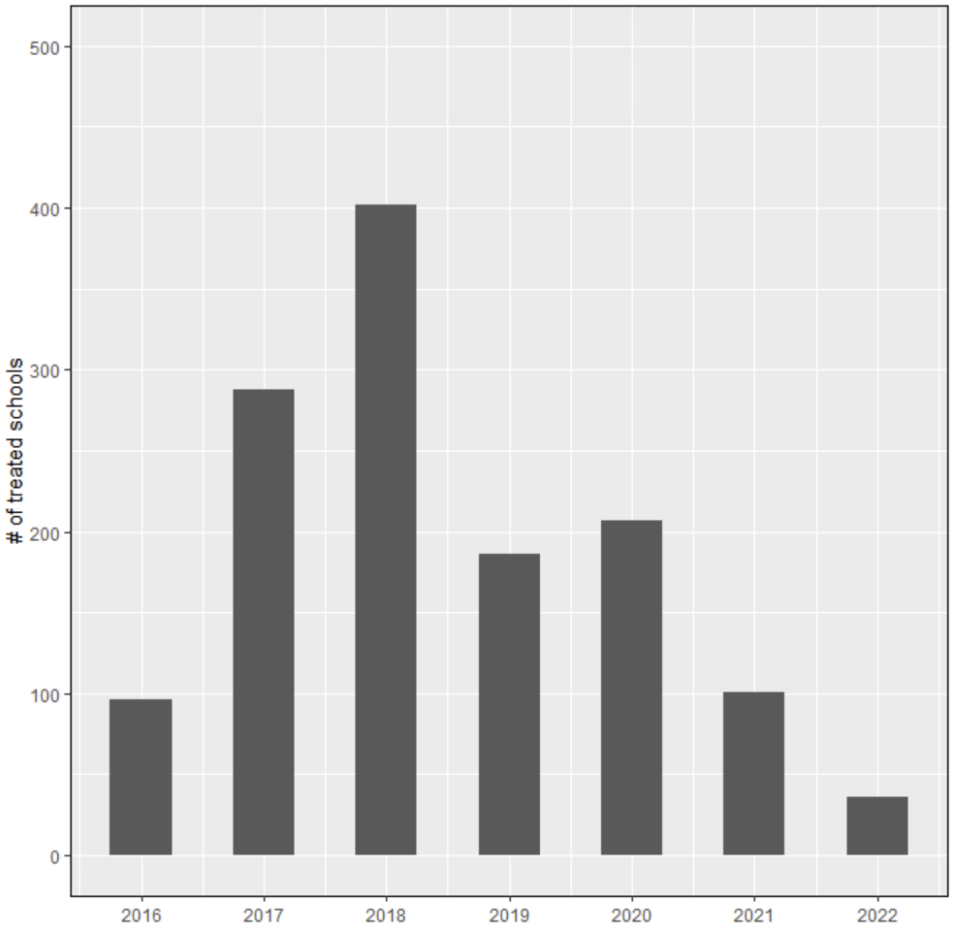
3.2 Treatment definition and SFL program characteristics

The SNAE tracked which schools received the SFL state grants in terms of the number of coaches assigned to each school. Based on these data, combined with the population register on the full population of students, teachers, and schools, we can identify which schools were treated and how many coaches were assigned per school. The main treatment definition identifies a school as being treated if they have at least one coach assigned to the school, as identified by the SNAE. As shown in Table 2, treated schools hosted on average approximately 1.7 coaches, 2.4 teacher groups, and about 16 participating teachers.

Based on this treatment definition, Figure 2 shows the staggered rollout in the total number of schools participating in SFL per year. The SFL dataset provided by the SNAE lists 1356 mainstream compulsory grade 1-9 schools which participated in SFL during the period studied, which corresponds to roughly one quarter of the schools in our data. One school was missing a school identification code and therefore could not be identified. Additionally, 108 schools were identified by SNAE as having received the treatment but no information on whether the school was assigned a coach could be found. We therefore exclude these schools so that the number of treated schools in our analysis is 1247. However, we test this alternative treatment definition in a sensitivity test. Unfortunately, we cannot test whether the effectiveness varied as a function of the proportion of participating teachers in a school, as these data are only available from the SNAE for a select number of years. However, schools which served grade 6 or 9 students had an average of 25 teachers in total working at a school based on teacher register data aggregated to the school level. If the average number of teachers per school (available from the SNAE for 2019-2022) is representative of participation rates for the other SFL program years, this indicates that on average roughly 65% of teachers in a school participated in SFL.

Figure 2

Participating grade 1-9 schools and rollout of SFL program



Note. The figure shows the number of schools participating in each wave of the SFL program between 2016 and 2022.

Table 2 shows the number of coaches per school, the average group size, the number of participating teachers, and the number of groups for participating mainstream grade 1-9 schools, based on data provided by the SNAE. When data for a variable were not available for all years, we based our calculations on the years with data and specify which years were used in the Table.

Table 2

SFL program characteristics of treated schools provided by the SNAE

	Data availability	Minimum	Maximum	Mean	SD
# of coaches	2016-2022	1	16	1.693	1.222
# of participants	2018-2022	1	86	16.45	11.47
# of groups	2018-2019	1	16	2.389	1.624
# teachers per group	2019-2022	1	40	9.996	5.664

Note. Participants refers to participating teachers.

3.3 Balancing tests and pre-tend assumptions

Table 3 presents a balancing test that compares the means of the pre-treatment treated schools as well as the never-treated schools prior to 2016. As expected in a large administrative dataset, several differences are statistically significant due to the large sample size and indicate slightly higher achievement, higher levels of parental education and lower levels of migrant and newly arrived students in the treated schools. However, the standardized magnitude of these differences was small. By contrast, the proportion of students with an individualized education program was virtually identical across treated and untreated schools.

For school-related variables, independent schools were substantially more common among treated schools than among never-treated schools. Treated schools were also more likely to be in the top achievement quartile and less likely to be in the bottom quartile of school achievement between 2014-2015. In addition, treated schools had slightly lower shares of certified teachers as well as slightly lower average years of teacher experience than never treated schools.

Table 3

Pre-intervention characteristics and balance tests of treated and never treated schools

	Never treated	Treated	*p*	*SMD*
Achievement measures				
Math (national test score)	-0.01	0.03	<0.001	0.042
Math (grade)	-0.01	0.03	<0.001	0.039
Swedish (national test score)	-0.01	0.03	<0.001	0.051
Swedish (grade)	-0.01	0.03	<0.001	0.049

Student characteristics				
Individualized education program	0.06	0.06	0.121	0.002
Parental education (1= post-secondary education)	0.60	0.62	<0.001	0.050
Migrant background	0.19	0.17	0.077	0.062
Childhood immigrant	0.11	0.10	0.010	0.092
Arrived within 5 years	0.07	0.06	0.003	0.111
School characteristics				
School type (1=independent school)	0.13	0.25	<0.001	0.302
Share certified teachers	0.74	0.72	<0.001	0.142
Share experienced teachers	13.46	13.11	0.004	0.099
Schools (*n*)	3689	1247		

Note. Parental education indicates the share of students whose parents have post-secondary degrees. Migrant background indicates they have two foreign born parents or are not born in Sweden. Both the national test scores and grades are standardized by each yearly cohort. SMD = standardized mean difference.

Although the patterns shown in Table 3 indicate that SFL-participating schools differed slightly from non-participating schools in several school and student-related characteristics, difference-in-differences estimates remain unbiased in the presence of such baseline differences, so long as the parallel trends assumption is met. This assumption implies that, absent SFL, the trajectory of grades and test scores in mathematics and Swedish would have been the same for treated and never-treated schools—even though their levels may differ. While there is no formal test for the parallel trends assumption, we assess its plausibility by examining pre-treatment trends across treated and never-treated schools for each outcome variable, using both visual inspection of the event study plots (Figure 3) and formal pre-trend tests (see supplementary materials section A). We find mostly stable pre-treatment estimates in the event-study plots, and the pre-trend tests in most cases do not reject the hypothesis that

treated and never-treated schools followed similar trends prior to treatment. There is an exception for mathematics national test scores, where the pre-trend test rejects the null hypothesis of no differential pre-trends, indicating that treated units were evolving differently from controls prior to treatment (see Table A1). However, the individual pre-treatment coefficients shown in the event study plots are very small in magnitude and not stable over time, suggesting that the rejection reflects minor changes rather than meaningful pre-trend differences. Moreover, the slightly positive pre-trend for treated schools in mathematics national test scores would, if anything, raise concerns about positive treatment effects for this outcome; however, no such effects are found, except for one subgroup effect for students in independent schools. No evidence of differential pre-trends is found for Swedish national test scores or for either Swedish or mathematics school grades.

3.4 Outcome measures

3.4.1 Grades and national test scores in mathematics and Swedish

To measure student achievement, we rely on two distinct measures: teacher-assigned grades in school years 6 and 9 and standardized national test scores in school years 6 and 9 in the subjects Swedish and mathematics. Teacher grades and national tests are scored at the scale F-A, where F is "fail." We translate this so that F is 0 and A is 5 and standardize both grades and student test scores (mean 0 and standard deviation 1) by year. We lack national test scores for 2020 and 2021 as no national tests were conducted in Sweden these years due to the COVID-19 pandemic.

We use both grades and national test scores as the outcome measures of interest for several reasons. In theory, teacher-assigned grades should reflect an overarching assessment of student general academic ability, based on continuous observation and evaluation over the academic year. They are also used for eligibility to upper-secondary education and from upper-

secondary school to tertiary education, making them highly consequential for students' further education (Vlachos, 2019). By contrast, national tests are provided at the end of an academic year and graded according to detailed instructions.

While national tests are not externally graded, schools are required to anonymize student responses before grading (SFS 2011:185, Chapter 9, Section 22) using nationally provided grading rubrics. Thus, national tests are less prone to grading leniency than grades. For our empirical strategy, permanent differences in grading leniency across schools would not cause bias. However, if treatment changes the extent of grading leniency, it would be a concern. We examine this by modeling grading leniency as an outcome in a sensitivity analysis.

4 Identification strategy

Due to the staggered rollout of SFL, approximately 25% of mainstream compulsory schools had participated in the SFL teacher professional development program by the end of the 2022/2023 academic year since its inception in 2016. We exploit this staggered rollout and test dynamic effects of the SFL program in grade 6 and 9 achievement using a difference-in-difference estimation strategy proposed by Callaway and Sant'Anna (2021). This approach addresses the concerns raised in recent evaluations of difference-in-differences analyses using two-way fixed-effects models (Roth et al., 2023) because it explicitly accommodates staggered treatment adoption and treatment effect heterogeneity over time, thereby avoiding bias which may arise when earlier-treated schools serve as inappropriate controls for later-treated schools and due to negative weighting of heterogeneous treatment effects. The Callaway and Sant'Anna (2021) estimator avoids these pitfalls by constructing separate cohort- and time-specific DiD comparisons rather than pooling all variation into a single regression coefficient.

The logic of this estimator can be illustrated by a scenario in which an intervention was allocated across 2 groups (treated and control) and 2 time periods (before and after). This can be expressed in the following, where D represents the DiD estimand:

$$D = (\mathbb{E}[Y_{post} | T = 1] - \mathbb{E}[Y_{pre} | T = 1]) - (\mathbb{E}[Y_{post} | T = 0] - \mathbb{E}[Y_{pre} | T = 0]) \quad (1)$$

The above formulation represents just two time periods (pre and post) and two groups (treated and untreated, T=1 or 0). However, in many cases, such as with the SFL program, there are multiple instances of pre- and post-treatment, many different units are treated, and treatment occurs at different times for different units. Moreover, there may be heterogeneous treatment effects across units. To accommodate staggered treatment timing and allow for heterogeneous effects across cohorts, Callaway and Sant'Anna (2021) propose a way to estimate average treatment effects on the treated (ATTs) for each treatment cohort *g* and each time *t*:

$$ATT_{g,t} = \mathbb{E}[Y_{st}(1) - Y_{st}(0) \mid G_s = g, t \geq g] \quad (2)$$

Where $ATT_{g,t}$ denotes the effect for cohort g at time *t*, $Y_{st}(1)$ is the average outcome for students in school *s* at time *t* after their school received the SFL intervention, and $Y_{st}(0)$ is the potential average student outcome under no treatment. This is approximated using never-treated schools as a control group, a choice that is supported by the large share of schools that never participated in SFL (approximately 75% of compulsory schools during the period in question), which ensures an informative and stable counterfactual group across all treatment cohorts. $G_s = g$ denotes that the school received treatment in year g, and $t \geq g$ restricts the analysis to the years after the school has received the treatment. Because we define treatment exposure based on whether a coach was assigned to the school, we interpret the estimates as intent-to-treat (ITT) effects, as schools assigned a coach may have varied in the degree to which the SFL program was implemented.

Identification of the average treatment effect rests on the parallel trends assumption: in the absence of SFL, the average student outcomes at treated schools would have followed the same trajectory as outcomes at schools that never participated in the program. The plausibility of this assumption is assessed empirically by examining pre-treatment estimates, which should be indistinguishable from zero if the assumption holds.

Put simply, the Callaway and Sant'Anna (2021) estimator recreates the logic of the canonical 2x2 DiD application by constructing cohort- and time-specific DiD comparisons that contrast treated units to an appropriate control group at each post-treatment period, and then aggregating these cohort-specific $ATT_{g,t}$ estimates into an overall treatment effect on the treated (ATT) weighted by the size of each treated cohort. Finally, we cluster standard errors at the school-level to account for within-school correlation in student outcomes.

4.1 Heterogeneity analyses

Given the broad mandate of the SFL program and its likelihood to disproportionately benefit several heterogeneous student groups, we conduct a number of heterogeneity analyses. First, we examine subgroups of students who are explicitly stated as the target of the SFL intervention: students with special educational needs (who have an individualized education program recorded in the national register) and students with a foreign language background (who either have two foreign born parents or who emigrated to Sweden at different times according to the national register).

We also conduct a series of more exploratory heterogeneity analyses in search of explanatory mechanisms for the findings, including gender, given that boys are more frequently identified as having special educational needs (Giota, Lace & Emanuelsson, 2023), school type, school achievement quartiles, and the average level of teacher qualifications in the school. We identify schools in the top and bottom quartiles of student achievement and average teacher

qualifications based on measures in the years prior to the SFL intervention (a two-year average based on 2014 and 2015).

4.2 Sensitivity analyses

As outlined in the previous section, we assess the plausibility of the parallel trends assumption (that treated and never-treated schools would have followed similar outcome trajectories in the absence of SFL) by testing for differential pre-treatment trends using an event study approach and a formal pre-trend test. Over and above this, we conduct several additional sensitivity checks to determine the robustness of our findings. Specifically, we investigate the degree to which our estimations of SFL program effects may be vulnerable to changes based on (1) treatment definition and model specification (in the latter, we test models with school type as a covariate and therefore change the identifying parallel trends assumption to conditional on school type), (2) student compositional changes, (3) program dosage (by investigating SFL effects in small schools), and (4) cohort-specific effects. We also investigate whether SFL influenced grading leniency by using grading leniency as the outcome in the difference-in-differences estimation.

5 Results

For each outcome measure, including mathematics and Swedish grades and national test scores, we present the results for the pooled sample. Next, we show results of the heterogeneity of SFL's effects for school subtypes and select student subgroups. Last, we present the results of the robustness and sensitivity tests.

5.1 Pooled analysis of SFL

As evidenced by the results shown in Table 4, we detect no clear pattern of effects across outcome measures when effects are averaged across all students. This is reflected by the very small effects in most cases, which are present for mathematics grades and national test scores as well as Swedish national test scores. We however find a small but statistically significant effect on students' Swedish grades, amounting to 0.023 standard deviations ($p = 0.026$).

Table 4

Estimated effects of SFL for all students

	Coefficient	Standard Error	z	p-value
Grades				
Math	0.0121	0.0089	1.36	0.173
Swedish	0.0228	0.0102	2.23	0.026
National test scores				
Math	0.0119	0.0162	0.74	0.462
Swedish	-0.0046	0.0242	-0.19	0.847

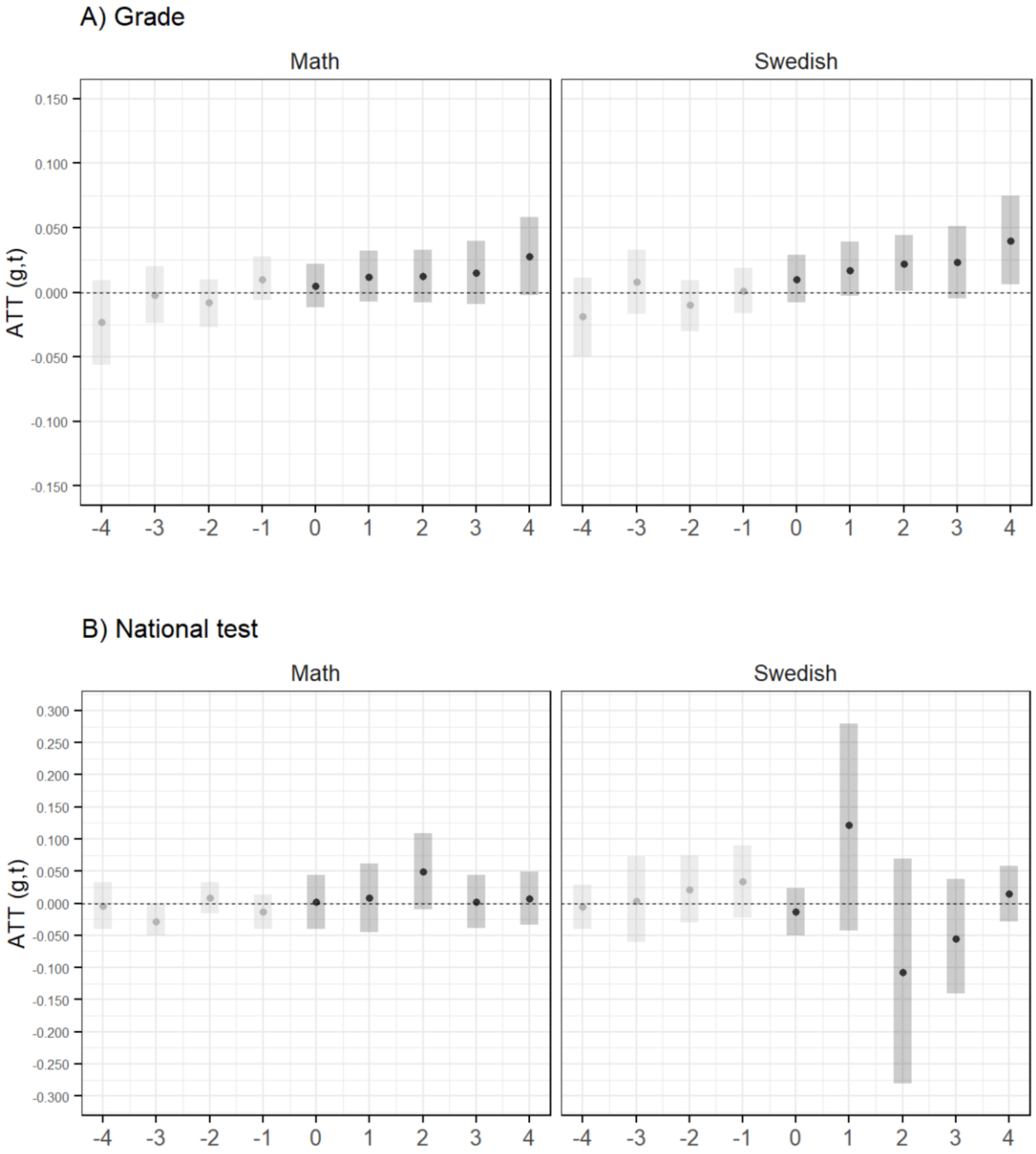
Note. The table shows the estimated effects of SFL on pooled Grade 6 and grade 9 to 8 math and Swedish grades and national test scores when pooled across all students. Scores have been standardized by outcome measure and year.

In Figure 3, we present the estimated dynamic effects of SFL by event time (the time periods in relation to the treatment) in both Swedish and mathematics by grades (panel A) and national test scores (panel B). As in Table 4, Figure 3 shows that for all outcomes except Swedish grades, the effect of SFL on mathematics and Swedish is negligible when averaged

across all students and grade 1-9 schools. Although the event-study plots show positive estimates after the treatment for both mathematics and Swedish grades which persist for several years, the post-SFL increases in grades are small, ranging from 0.01-0.04 standard deviations, and in most cases not reaching statistical significance. This pattern, however, is not reflected in the national test scores, which show consistently null estimates in the case of mathematics and very unstable estimates in the case of Swedish.

Figure 3

Event-time SFL effects on all students in grade 1-9 schools



Note. The figure displays event-time estimates where period 0 on the x-axis denotes the first year of SFL participation. Negative x-axis values indicate years prior to treatment and positive values indicate years after treatment. Light grey markers correspond to pre-treatment periods and dark grey markers correspond to post-treatment periods (point estimates and 95% confidence intervals). ATT(g,t) = group time average treatment effect on the treated (Callaway & Sant'Anna, 2021).

5.2 Heterogeneity in the effects of SFL

We turn to reporting the effects of SFL program participation for subgroups of students and different school profiles. We begin by subsetting the data by student subgroups, including students with individualized education programs, students with two foreign born parents, childhood immigrants, and newly arrived students. Next, we re-estimate the difference-in-difference estimates for students by school achievement quartile and average teacher qualifications in the school, as well as by school type.

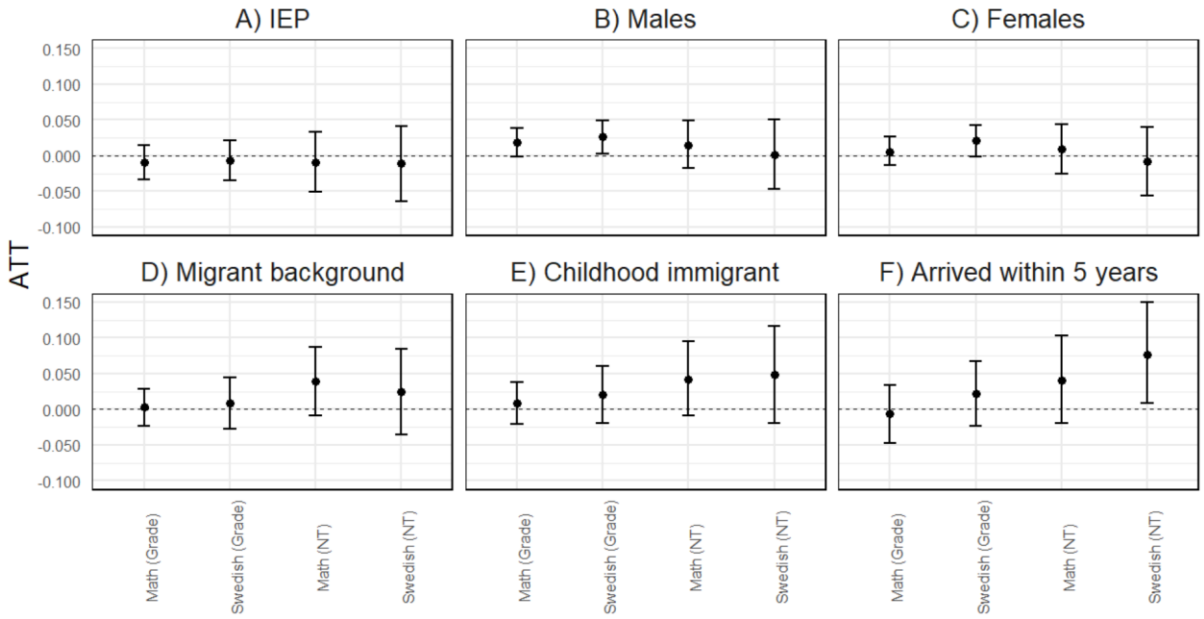
We first present the results for students who have had individualized educational programs at any point in their educational history in Figure 4. Figure 4 shows average effects of -0.009 and -0.007 standard deviations for mathematics and Swedish grades, respectively, and -0.01 and -0.016 standard deviations for mathematics and Swedish national test scores, indicating null or slightly negative effects of SFL for students with special educational needs. Given the focus of the SFL program, this result is certainly noteworthy. Figure 4 also shows that effects of SFL did not systematically differ by gender for the most part, except in the case of mathematics grades where effects were larger for boys than for girls.

Last, Figure 4 shows the average effects for students who have foreign language backgrounds, either due to two foreign born parents or who were not born in Sweden. We measure the latter separately for childhood immigrants (arrived any time after birth), and students arriving within 5 years for each cohort. For students who arrived within 5 years, we find positive effects in Swedish and mathematics national tests of 0.077 (p = 0.047) and 0.04 standard deviations, respectively, though only in the former case does the effect reach (marginal) statistical significance. Additionally, these effects do not generalize to the teacher-assigned grades for mathematics nor Swedish. This pattern is similar for childhood immigrants, though the national test scores are increased by roughly 0.05 standard deviations and they are not statistically significant. We find less convincing evidence of this pattern when extended to

include those with two foreign born parents. Event study estimates for student subgroups are shown in supplementary material section B.

Figure 4

SFL effect estimates by student characteristics



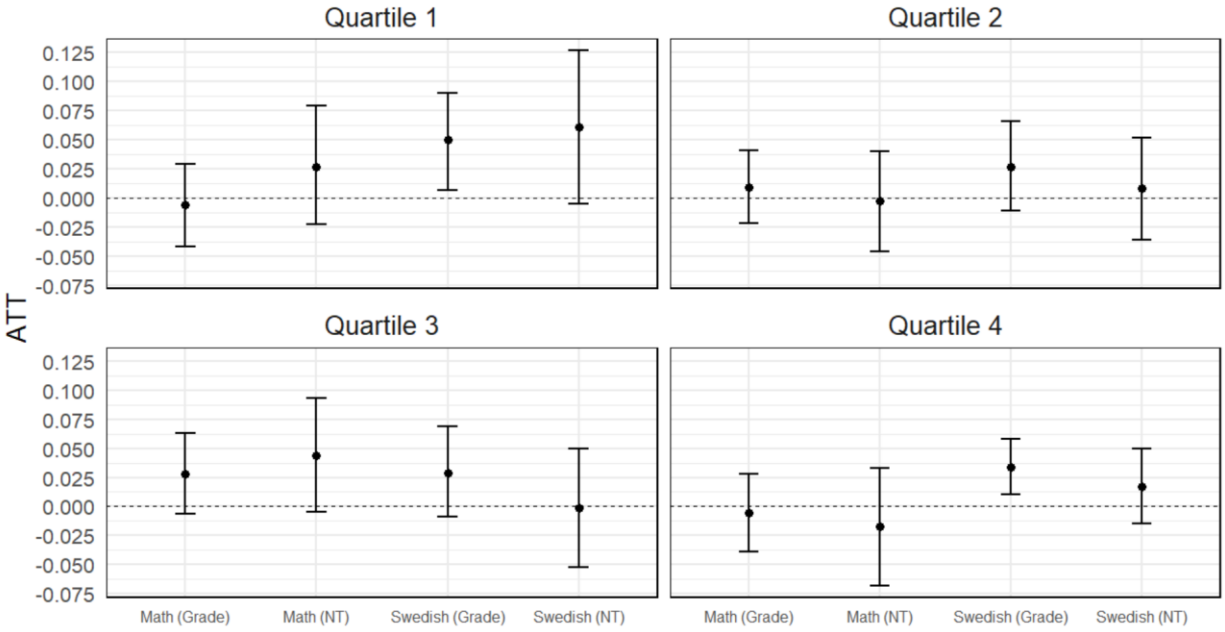
Note. Point estimates and 95% confidence intervals are displayed by student subgroup (IEP, gender, migration background). IEP=individualized education program. Migrant background = has two foreign born parents. Childhood immigrant = student born outside of Sweden. Arrived within 5 years = migrated to Sweden within 5 years. ATT = average treatment effect on the treated.

Figure 5 shows average effects by school achievement quartiles prior to the introduction of the SFL program. The school achievement quartiles are determined based on the subject in focus. For example, if the outcome measure is mathematics grades we use average school achievement in mathematics grades prior to the intervention to determine schools in the top and bottom quartile. Figure 5 shows a null pattern of effects for both mathematics and Swedish, with some exceptions. Most notably, we find positive and statistically significant effects of the SFL intervention on student grades in Swedish for those in the bottom achievement quartile, amounting to about 0.05 standard deviations for grades (*p*

= 0.022) and 0.06 standard deviations for national test scores though not statistically significant. Schools in the top achievement quartile also show small positive effects for Swedish grades (0.035 standard deviations, $p < 0.005$), but the effects in Swedish national test scores do not reach statistical significance. We show event study plots for top and bottom quartile schools in the supplementary material section C.

Figure 5

SFL effect estimates by school achievement quartile



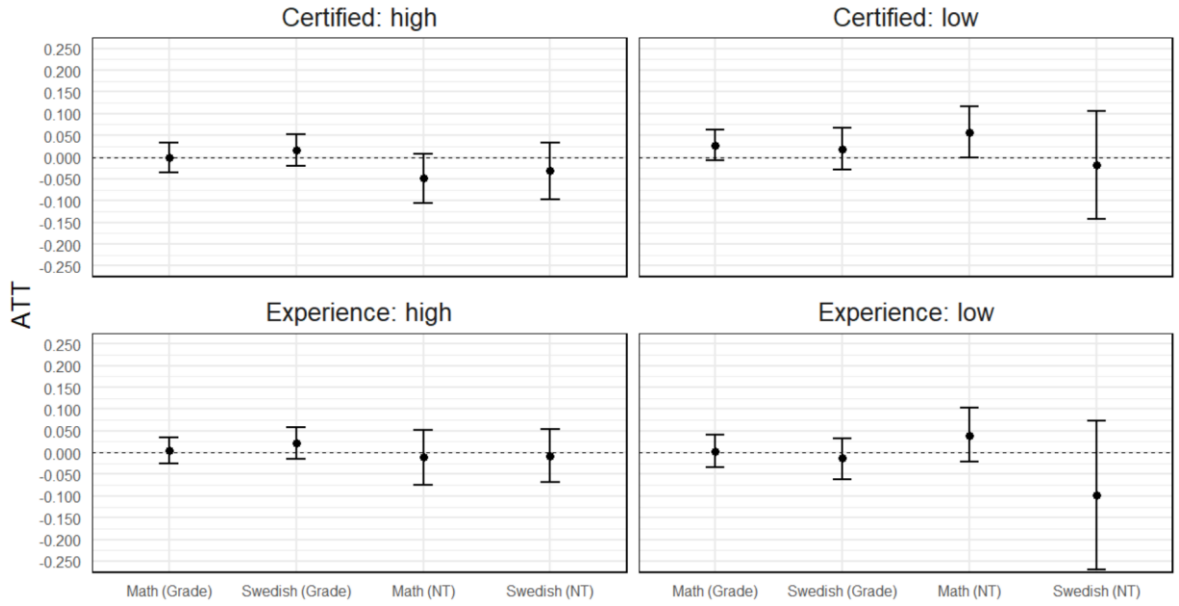
Note. Figure shows effects of SFL by school achievement quartile. Quartile 1 refers to the bottom achievement quartile and quartile 4 refers to the top achievement quartile (measured as a 2-year average in school achievement before the SFL intervention). ATT=average treatment effect on the treated.

We find a similar pattern of null effects when subsetting the sample by average teacher qualifications, either in the proportion of certified teachers or by average years of experience, as shown in Figure 6. One exception to this is a marginally insignificant treatment effect in mathematics national test scores for students in schools with a low share of certified

teachers of 0.057 standard deviations ($p = 0.052$). In all other cases, effects of SFL do not differ by the average teacher qualification level in a school.

Figure 6

SFL effect estimates by average teacher qualifications

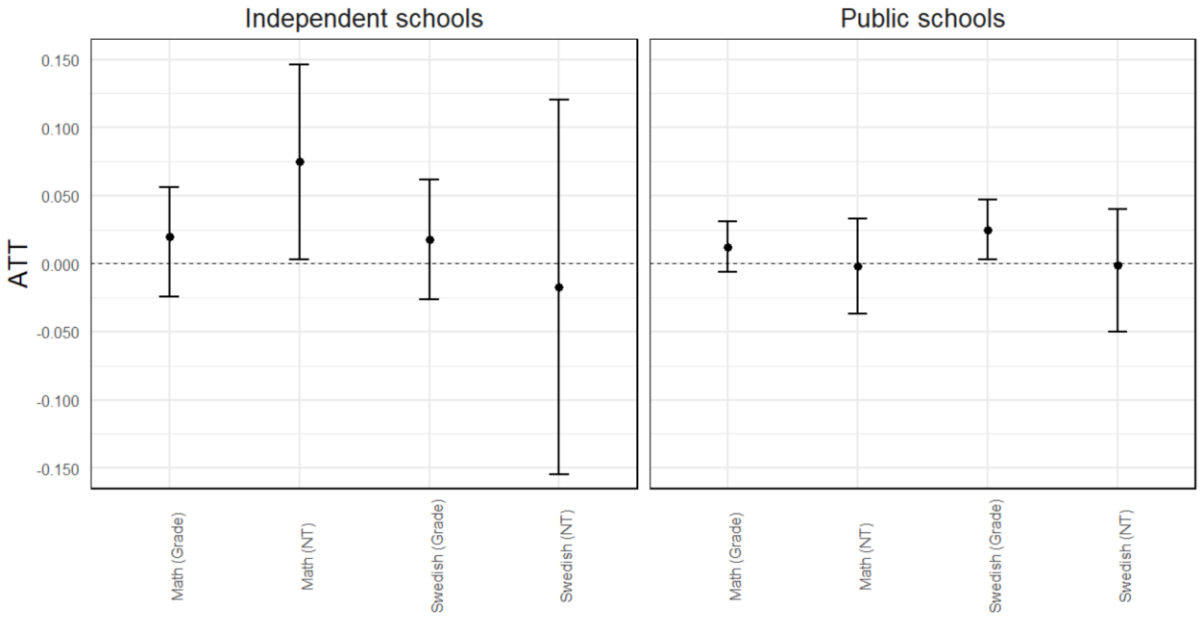


Note. Figure shows effects of SFL by schoolteacher qualification level. Low refers to schools in the bottom quartile of the distribution of the share of certified teachers or average level of teacher experience, and high refers to schools in the top quartile of the certification/ experience distribution before the SFL intervention. ATT=average treatment effect on the treated.

Finally, we do not find any differential effects when analyses are subset by school type (see Figure 7), save for a similar pattern of positive effects in mathematics national test scores for students in independent schools (0.075 standard deviations, $p = 0.04$). The similarity of these estimates with those for schools with a low share of certified teachers may be due to that independent schools employ uncertified teachers at a higher rate. Event study plots for analyses subset by school type can be found in supplementary materials section C.

Figure 7

SFL effect estimates by school type



Note. Figure shows effects of SFL by school type. ATT=average treatment effect on the treated.

5.3 Sensitivity analyses

We report the results from several sensitivity checks to determine the robustness of our findings in the study supplementary materials. First, we consider an alternative treatment specification where we include all schools identified by the SNAE as having participated in the SFL program despite the lack of information on whether a coach was assigned to the school. The event study plots and ATTs for this analysis can be found in supplementary material section D, showing average pooled estimates highly similar to those we find in the beginning of this results section in Table 4, indicating general stability of findings across treatment specifications. We also test the alternative treatment specification for the statistically significant findings from the heterogeneity analyses and find effects of similar magnitude but which in all cases fail to reach conventional levels of statistical significance.

Next, although we do not have access to implementation data across each year, we examine whether treatment dosage has an impact on SFL success by investigating whether small schools showed larger effects of SFL (based on the assumption that participating teacher

shares will be higher in small schools). These results can be found in supplementary material section E, showing slightly larger effects in mathematics though they do not reach statistical significance.

To assess whether observed treatment effects might reflect changes in the demographic and socioeconomic composition of schools' student bodies rather than genuine program effects, we follow Grönqvist et al. (2025) in examining whether the demographic and socioeconomic composition of schools changes systematically following SFL implementation. We operationalize student body composition using students' predicted test scores (a composite measure constructed from students' demographic and socioeconomic characteristics, including student migration background, gender, parental education, and school fixed effects) and test whether this measure shifts around the time of SFL adoption. These estimates can be found in section F of the supplementary materials and show no indication of changes in student composition which could serve as an alternative explanation. Next, we examine the concern of grading leniency as an SFL outcome in section G of the supplementary materials. Effects are negligible for mathematics, while effects for Swedish are positive but statistically insignificant at approximately 0.02 standard deviations. The latter suggests that increased grading leniency remains a plausible explanation for the observed improvements in Swedish grades, particularly given the absence of corresponding effects on national test scores.

Given that earlier participating schools may differ systematically from schools which participated in later SFL cohorts, and to account for any minor changes to program design and content over time, we also examine cohort specific effects to examine whether SFL effectiveness varied as a function of the cohort of participating schools, and these estimates can be found in supplementary materials section H, and show that the null effects of SFL cannot be explained by effect heterogeneity across cohorts.

Finally, considering that independent schools were more likely to participate in the SFL program, we incorporate school type as a covariate in the estimation and therefore change the identifying parallel trends assumption to conditional on school type. The results are presented in supplementary materials section I, showing nearly identical estimates to those in the main specification.

6 Discussion and conclusion

Through a difference-and-differences estimation strategy, this study evaluated the effects of a large-scale teacher professional development program in Sweden aimed at training general classroom teachers in special and inclusive education—a program in which 60,000 teachers participated and which cost the equivalent of roughly 70 million USD. In relation to previous research on PD programs in special and inclusive education for ordinary classroom teachers, our analysis constitutes rare and robust evidence on the dynamic effects of such programs—evidence which is particularly scarce for large-scale programs such as SFL (Donath et al., 2023).

The effects of the SFL program are relevant to a global educational policymaking audience for several reasons. Like most countries, Sweden has a diverse body of pupils, ranging from students with special educational needs, foreign language backgrounds, or who may otherwise struggle in school, the lion's share of whom attend general classrooms. General classroom teachers repeatedly request additional training in this area (Cooc, 2019; Eurydice, 2019). Causal evidence on in-service initiatives aimed at improving teaching in general classrooms can therefore tell us which approaches are successful and whether their continued or wider use is justified. Moreover, using a dynamic difference-in-differences approach can tell us if effects of such programs persist in the post-intervention period.

Overall, the story regarding the success of the SFL program is mixed. When averaged across all students, SFL yielded small and statistically significant positive effects in Swedish grades of 0.023 standard deviations, persistent for several years after the intervention. However, the average effects of SFL were not confirmed by the national test scores, which indicates that grade improvements may be driven by changes in grading leniency, a result which was formally tested and indicated in the case of Swedish (though not reaching statistical

significance). We are therefore unable to conclude that the program was successful in raising student achievement for the general student body. However, given the specific focus of SFL in targeting (1) students with individualized education programs, and (2) students with foreign language backgrounds, we argue that the effect heterogeneity is more relevant for determining the success of SFL.

For students with individualized educational programs, we find consistently null or even slightly negative effects of the SFL program. This finding is indeed discouraging given the explicit focus of the SFL program on special educational approaches. On the other hand, a subset of childhood immigrants may have benefited from SFL—specifically those arriving in Sweden within 5 years. Students attending SFL participating schools in this group showed improved test scores of 0.077 standard deviations in national test scores in Swedish, a magnitude that is often described as moderate relative to effects commonly observed in educational interventions (Evans & Yuan, 2020; Kraft, 2020). These effects occurred for the most part 1-2 years after the first treatment year but were not stable in the years following despite some positive effects being found in later years. In addition, the estimates reflect the impact of offering SFL at the school level. Considering that not all teachers in participating schools took part in SFL, the reported effects may underestimate any improvement that might have been observed under full participation. At the same time, these findings should be interpreted cautiously as they stem from subgroup analyses.

Due to the null effects on students with individualized education programs, we conclude that our analysis does not add empirical support for the type of broad PD focusing on whole-class approaches for students with special educational needs used in the SFL program, which included a host of diverse aims, for example including improved student-teacher relationships or a more safe and supportive school climate, via enhanced teacher awareness, reflection and practice with respect to a wide range of (typically inclusion-oriented) topics in

special education. This may be due to either the PD design or the pedagogical approaches advocated by the program. If the latter is the case, the results are consistent with Fuchs et al. (2025), who argue that research offers inconclusive evidence regarding how to best support students with disabilities in general classrooms, while the strongest and most reliable findings instead point to the effectiveness of intensive supplementary instruction, typically delivered outside the general classroom. An alternative interpretation is that there are more effective practices than those promoted within the SFL program (Mccrea et al., 2025). Indeed, while the findings are more scarce and the evidence more mixed, studies have documented whole-class programs that improve outcomes for struggling students (Neitzel et al., 2022).

The results however suggest that the program was beneficial for newly arrived immigrant students. Given the scarce body of evidence documenting the impact of inclusive and special educational approaches for this population, our findings are difficult to contextualize within a wider frame. In addition, though newly arrived students may benefit to a greater degree from the pedagogical strategies promoted within the SFL program, it is also possible that schools serving different student bodies implemented different modules of SFL. For example, the positive effects for migrant students could be due to schools serving larger shares of these students implementing the multilingualism module.

Though we do not have insight into the classroom practices which may have led to the documented effects, we investigate results across a range of school categories and student characteristics in search of explanatory mechanisms for our findings. First, we observe positive effects in Swedish across both achievement measures in schools which were in the bottom achievement quartile prior to the SFL intervention, though only reaching statistical significance for grades. This finding mirrors those from the previous Swedish large-scale PD program BFM, which similarly were not found to raise test scores for low achieving students (Grönqvist et al., 2025).

Second, a persistent pattern of null effects in mathematics was evident throughout the results. As we have noted, one exception to this finding is a positive effect of 0.075 standard deviations in mathematics national test scores in independent schools and 0.06 in schools with a high share of uncertified teachers, possibly due to independent schools employing a higher share of uncertified teachers. This may be due to increased sensitivity of uncertified teachers to in-service training, which was also found the evaluation of the BFR program (Holmlund et al., 2024). However, these estimates were reduced considerably in magnitude and no longer statistically significant under model specifications with an alternative treatment definition.

The null or weak effects of the SFL program may not necessarily be due to the special education approaches advocated by the program. It may, alternatively, be the form of the program that reduced its success. Like previous large-scale PD programs in Sweden, the SFL program followed a 'collegial learning' approach, where teachers follow modules and meet weekly to discuss module content, apply it in their classrooms, and reflect on the classroom application together under the guidance of a trained coach. Although the program included some effective PD elements (e.g., coaching and theory-practice links, Kraft et al., 2018; Sims et al., 2023), it placed relatively little emphasis on deliberate modelling, rehearsal, and feedback focused on specific instructional practices (Banks et al., 2025; Cohen et al., 2024; Sims et al., 2026). This was not an accidental omission: the program deliberately avoided promoting specific teaching methods, instead offering materials intended to stimulate professional dialogue within schools about improvement. While this approach may have merits, it may also have limited the program's impact on classroom practice. At the same time, evidence from two earlier Swedish programs using a similar 'collegial learning' model (BFR and BFM) suggests that such interventions tend to yield only small effects (0.01 to 0.03 standard deviations in national test scores), yet these programs showed positive impacts

nonetheless. The absence of any detectable effects for SFL therefore points to differences in program content as a possible explanation for the different results across programs.

Future policy efforts to boost student test scores may nevertheless wish to refine certain elements of the Swedish 'collegial learning' approach. For example, considering that a long duration has not been shown to be important for PD success (Didion et al., 2020; Kraft et al., 2018; Lynch et al., 2025) and that teachers' time is limited, a shorter and narrower approach may be warranted. Furthermore, given the inherent challenges in scaling programs while maintaining effects, it would be valuable to integrate rigorous evaluation into program rollout from the outset, enabling A/B comparisons of design alternatives and evidence-based refinements before nationwide implementation (Angrist et al., 2026; List, 2024).

A final noteworthy finding in our study concerns the positive effects of SFL for newly arrived childhood immigrants which were found in the national tests and not teacher-assigned grades, a finding in contrast to the pooled results, where effects were found in grades but not in the national tests. The fact that national tests, unlike grades, do not capture all forms of knowledge covered in the curriculum means that some differential effects across these outcomes may arise. For example, some groups of students may perform well on the national tests but be less engaged with their coursework over the school year. Despite this, given that the SFL program was partly intended to increase the share of students with a foreign language background eligible for upper-secondary education, (an outcome determined by school grades) its effectiveness in this regard appears to have been limited (SNAE, 2025).

The study has some limitations. The main limitations have to do with opacity regarding SFL program implementation, as beyond the foundational mandatory modules on inclusion, we cannot say for certain which PD modules were implemented in which schools and to what degree the modules were followed. We however argue that this presents a trade-

off ultimately outweighed by the ecological validity of the findings due to the large scale and detailed analyses across subgroups and time. Though the estimates undoubtedly reflect considerable implementation heterogeneity, we partially address this limitation by examining differential impacts based on school size, teacher qualifications, and school type. In addition, it is possible that the SFL program may have been beneficial for certain subgroups of special educational needs, but we are not able to examine this based on the national register data. Furthermore, it is also possible that the program might have produced effects on other outcome measures than those tested in this study, such as more narrow dimensions of reading or mathematics, or behavioural outcomes. Finally, although the parallel trends assumption held in most cases and we found no effects of the SFL program on predicted test scores, participation in the SFL program was voluntary; as a result, participating schools may differ in unobserved characteristics, potentially violating the parallel trends assumption, which suggests that the certainty of causal interpretations are weaker than under randomized assignment to program participation.

Despite these limitations, in light of the still limited causal evidence on large-scale PD in special or inclusive educational approaches, our study is indeed useful. Teacher PD programs remain a critical lever for improving teaching practices and student academic outcomes, requiring substantial financial and time commitments on the part of school actors and school staff. Evaluating their performance, therefore, is critical to justify further investment and program deployment. This is particularly true for special and inclusive education initiatives, where high-quality evidence remains scarce despite that more PD for general education teachers in this area is needed (Cooc, 2019; Didion et al., 2026; Eurydice, 2019). Our results show that one such large-scale special education PD program in Sweden did not achieve its intended aim of improving the academic outcomes of all students or students with

individualized education programs, though it was partially successful in boosting the Swedish outcomes of newly arrived immigrant students.

References

Angrist, N., et al. (2026). Iterative A/B testing for social impact: rigorous, rapid, regular. *Stanford Social Innovation Review,* https://ssir.org/articles/entry/iterative-a-b-testing-social-impact

Bal, A., & Perzigian, A. B. T. (2013). Evidence-based Interventions for Immigrant Students Experiencing Behavioral and Academic Problems: A Systematic Review of the Literature. *Education and Treatment of Children*, *36*(4), 5–28. http://www.jstor.org/stable/42900224

Banks, B., Sims, S., Curran, J., Meliss, S., Chowdhury, N., Altunbas, H. G., … Instone, I. (2025). Decomposition and recomposition: effects on early-career teachers' enactment and adaptive transfer of behaviour management practices. *European Journal of Teacher Education*, 1–24. https://doi.org/10.1080/02619768.2025.2511867

Bölte, S., Leifler, E., Berggren, S., & Borg, A. (2021). Inclusive practice for students with neurodevelopmental disorders in Sweden. *Scandinavian Journal of Child and Adolescent Psychiatry and Psychology, 9,* 9–15. https://doi.org/10.21307/sjcapp-2021-002

Callaway, B., & Pedro H.C. Sant'Anna, P. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics, 225*, 200-230, https://doi.org/10.1016/j.jeconom.2020.12.001.

Carroll, J. M., Bradley, L., Crawford, H., Hannant, P., Johnson, H., & Thompson, A. (2017). *SEN support: A rapid evidence assessment*. Coventry University.

Chao, C. N. G., Sze, W., Chow, E., Forlin, C., & Ho, F. C. (2017). Improving teachers' self-efficacy in applying teaching and learning strategies and classroom management to students with special education needs in Hong Kong. *Teaching and Teacher Education,* 66, https://doi.org/10.1016/j.tate.2017.05.004.

Chen, X., & Zhang, Y. (2019). Typical practices of lesson study in East Asia. *European Journal of Education*, 54, 189–201. https://doi.org/10.1111/ejed.12334.

Choi, S., Mao, X., & Park, S. (2025). Promoting school belonging for immigrant students: the interplay of inclusive school climate and multicultural education. *School Effectiveness and School Improvement*, *36*(4), 522–542. https://doi.org/10.1080/09243453.2025.2492012

Choi, S., & Lee, S. W. (2020). Enhancing Teacher Self-Efficacy in Multicultural Classrooms and School Climate: The Role of Professional Development in Multicultural Education in the United States and South Korea. *AERA Open*, *6*(4). https://doi.org/10.1177/2332858420973

Coffey, S. G., Goodman, J., Schwartz, A. E., Stiefel, L., Winters, M. A., & Yoon, Y. H. (2026). Special education substantially improves learning: Evidence from three states. NBER Working Paper No. 34998. https://doi.org/10.3386/w34998

Cohen, J., Wong, V. C., Krishnamachari, A., & Erickson, S. (2024). Experimental Evidence on the Robustness of Coaching Supports in Teacher Education. *Educational Researcher*, *53*(1), 19-35.

Cooc, N. (2019). Teaching students with special needs: International trends in school capacity and the need for teacher professional development. *Teaching and Teacher Education, 83*, 27–41. https://doi.org/10.1016/j.tate.2019.03.021

Dee, T.S., & Jacob, B. (2011), The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30,* 418-446. https://doi.org/10.1002/pam.20586

Deunk, M. I., Smale-Jacobse, A. E., de Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation practices: A systematic review and meta-analysis of studies on the cognitive effects of differentiation practices in primary education. *Educational Research Review, 24*, 31–54.

Didion, L., Toste, J. R., & Filderman, M. J. (2020). Teacher professional development and student reading achievement: A meta-analytic review of the effects. *Journal of Research on Educational Effectiveness, 13*(1), 29–66. https://doi.org/10.1080/19345747.2019.1670884

Didion, L., Filderman, M. J., Hart, E. J., Payne, S. B., Vitale, J., Olmstead, C., & Wexler, J. (2026). High-Quality Special Education Professional Development Improves Teacher and Student Outcomes. *The Journal of Special Education* https://doi.org/10.1177/00224669261423924

Donath, L., Luke, T., Graf, E., Tran, U., Götz, T. (2023). Does professional development effectively support the implementation of inclusive education? A meta-analysis. Educational Psychology Review, 35, https://doi.org/10.1007/s10648-023-09752-2

Eurydice. (2019). Integrating Students from Migrant Backgrounds into Schools in Europe: National Policies and Measures. Eurydice Report. Luxembourg: Publications Office of the European Union.

Expressen. (2023, April 29). *Skolverket sprider myter och duckar för kritiken*. Expressen. https://www.expressen.se/debatt/skolverket-sprider-myter-och-duckar-for-kritiken/

Evans, D. K., & Yuan, F. (2020). How big are effect sizes in international education studies? *Educational Evaluation and Policy Analysis.* 01623737221079646.

Florian, L. (2019). On the necessary co-existence of special and inclusive education. *International Journal of Inclusive Education, 23,* 7-8, 691-704, DOI: 10.1080/13603116.2019.1622801

Fuchs, L. S., Fuchs, D., Compton, D. L., Wehby, J., Schumacher, R. F., Gersten, R., & Jordan, N. C. (2015). Inclusion Versus Specialized Intervention for Very-Low-Performing Students: What Does Access Mean in an Era of Academic Challenge? What Does Access Mean in an Era of Academic Challenge? *Exceptional Children*, *81*(2), 134-157.

Fuchs, L. S., Sterba, S. K., Fuchs, D., & Malone, A. S. (2016). Does Evidence-Based Fractions Intervention Address the Needs of Very Low-Performing Students? *Journal of Research on Educational Effectiveness*, *9*(4), 662–677. https://doi.org/10.1080/19345747.2015.1123336

Fuchs, D., Gilmour, A. F., & Wanzek, J. (2025). Reframing the most important special education policy debate in 50 years: How versus where to educate students with disabilities in

America's schools. *Journal of Learning Disabilities*, *58*(4), 257-273. https://doi.org/10.1177/00222194251315196

Gamse, B. C., Bloom, H. S., Kemple, J. J., & Jacob, R. T. (2008). *Reading First impact study: Interim report* (NCEE 2008-4016). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Gersten, R., Haymon, K., Newman-Gonchar, R., Dimino, Jayanthi, M. (2020). Meta-analysis of the impact of reading interventions for students in the primary grades. *Journal of Research on Educational Effectiveness, 13*, https://doi.org/10.1080/19345747.2019.1689591.

Giota, J., Lace, I., & Emanuelsson, I. (2023). School achievement and changes in inclusive *vs* exclusive support over 50 years in Sweden regarding students with intellectual disabilities and special educational needs. *Scandinavian Journal of Educational Research*, *67*(6), 997–1011. https://doi.org/10.1080/00313831.2022.2115129

Grönqvist, E., Öckert, B., & Rosenqvist, O. (2025). Does the "Boost for Mathematics" Boost Mathematics? A Large-Scale Evaluation of the "Lesson Study" Methodology on Student Performance. *American Economic Journal: Economic Policy, 17,* 345-372.

Holmstrand, L. & Härnsten, G. (2003). Förutsättningar för forskningscirklar i skolan: En kritisk granskning. Stockholm: Myndigheten för skolutveckling.

Holmlund, H., Häggblom, J., & Lindahl, E. (2024). The boost for reading. IFAU-Institute for Evaluation of Labour Market and Education Policy. IFAU Report 2024:5.

Jacob, B., & Lefgren, L. (2004). The impact of teacher training on student achievement quasi-experimental evidence from school reform efforts in Chicago. *The Journal of Human Resources, 39,* 50–79.

Jordan, A., Schwartz, E., & McGhie-Richmond, D. (2009). Preparing teachers for inclusive classrooms. *Teaching and Teacher Education, 25*(4), 535–542. https://doi.org/10.1016/j.tate.2009.02.010

Kahmann, R., Droop, M., & Lazonder, A.W. (2022). Meta-analysis of professional development programs in differentiated instruction. *International Journal of Educational Research, 116*, 1-13, 10.1016/j.ijer.2022.102072

Kirsten, N., & Carlbaum, S. (2020). Kompetensutveckling för professionella lärare? Introduktionen av kollegialt lärande i svensk skola. *Pedagogisk forskning i Sverige,* DOI:

https://doi.org/10.15626/pfs25.01.01

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241-253. https://doi.org/10.3102/0013189X20912798

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. Review of Educational Research 88 (4): 547–88. https://doi.org/10.3102/0034654318759268.

List, J.A. (2024). Optimally generate policy-based evidence before scaling. *Nature, 626,* 491-499.

Lynch, K., Gonzalez, K., Hill, H., & Merritt, R. (2025). A meta-analysis of the experimental evidence linking mathematics and science professional development interventions to teacher knowledge, classroom instruction and student achievement. AERA Open, https://doi.org/10.1177/23328584251335302.

Machin, S., & McNally, S. (2008). The literacy hour. *Journal of Public Economics, 92*, 1441–1462.

Mccrea, P., Goodrich, J., & Barker, J. (2025). Inclusive teaching – a discussion paper. Inclusive Teaching: A New Approach for SEND Challenges

Neitzel, A. J., Lake, C., Pellegrini, M., & Slavin, R. E. (2022). A synthesis of quantitative research on programs for struggling readers in elementary schools. *Reading Research Quarterly, 57*(1), 149–179.

OECD. (2015a). *Improving Schools in Sweden: An OECD Perspective*. Paris: OECD publishing.

OECD (2015b), *Immigrant Students at School: Easing the Journey towards Integration*, OECD Reviews of Migrant Education, OECD Publishing, Paris, https://doi.org/10.1787/9789264249509-en.

OECD. (2021). Promoting inclusive education for diverse societies: A conceptual framework. OECD Working Papers. EDU/WKP(2021)17.

O'Hagan, K.G., & Stiefel, L. (2024). Does special education work? A systematic literature review of evidence from administrative data. *Remedial and Special Education, 46,* https://doi.org/10.1177/07419325241244485

Pang, M. F., & Ling, L. M. (2012). Learning study: Helping teachers to use theory, develop professionally, and produce new knowledge to be shared. *Instructional Science, 40*(3), 589–606. https://doi.org/10.1007/s11251-011-9191-4

Regeringen. (2015). *Uppdrag att svara för genomförandet av fortbildning i specialpedagogik för lärare i grundskolan, motsvarande utbildning vid särskilda ungdomshem och sameskolan* (Regeringsbeslut I:3, U2015/05783/S). Utbildningsdepartementet.

Savolainen, H., Engelbrecht, P., Nel, M.,& Malinen, O.- P. (2012). Understanding teachers' attitudes and self-efficacy in inclusive education: Implications for pre-service and in-service teacher education. *European Journal of Special Needs Education, 27,* 51–68.

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,* 480-520. https://doi.org/10.1002/pam.22282

Scanlon, D., Gelzheiser, L.M., Vellutino, F., Schatschneider, C., Sweeney, J.M. (2008). Reducing the incidence of early reading difficulties: Professional Development for classroom teachers versus direct interventions for children, *Learning and Individual Differences, 18,* 346-359, https://doi.org/10.1016/j.lindif.2008.05.002.

Sims, S., Fletcher-Wood, H., Godfrey-Faussett, T., Mccrea, P., & Meliss, M. (2026). Modelling evidence-based practice in initial teacher training: effects on teachers' skills,

knowledge and self-efficacy. *Instructional Science, 54*, https://doi.org/10.1007/s11251-026-09779-2.

SFS 2011:185. *Skolförordning* [Education Ordinance]. Utbildningsdepartementet. https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/skolforordning-2011185_sfs-2011-185/

Söderlund, G., Thoutenhoofd, E., Westman Andersson, G., Gerrbo, I., Berhanu, G. (2024). Inclusive education in Sweden: a survey of teachers' attitudes, self-efficacy and collaboration towards better meeting pupils learning. *European Journal of Inclusive Education, 3,* https://doi.org/10.7146/ejie.v3i1.145185

Swedish National Agency for Education (SNAE). (2016). Redovisning av regeringsuppdrag Dnr 2016:20. https://www.skolverket.se/publikationer?id=3986

Swedish National Agency for Education (SNAE). (2025). *Specialpedagogik för lärande: Slutredovisning av regeringsuppdrag om fortbildning i specialpedagogik 2015–2025* (Skolverket rapport 2025:2853). Skolverket.

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Goodrich, J., Van Herwegen, J., & Anders, J. (2023). Effective Teacher Professional Development: New Theory and a Meta-Analytic Test. *Review of Educational Research*, *95*(2), 213-254. https://doi.org/10.3102/00346543231217480

Smale-Jacobse, A.E., Meijer, A., Helms-Lorenz, M., Maulana, R. (2019). Differentiated instruction in secondary education: A systematic review of research evidence, Frontiers in Psychology, 10, 1-23, 10.3389/fpsyg.2019.02366

Thoutenhoofd, E. D., Söderlund, G., Westman Andersson, G., Berhanu, G., & Gerrbo, I. (2020). *Utvärdering av kompetensutvecklingsinsatsen specialpedagogik för lärande (SFL).* Institutionen för pedagogik och specialpedagogik, Göteborgs universitet. (RIPS: Rapporter från Institutionen för pedagogik och specialpedagogik, nr 20).

Vlachos, J. (2019). Trust-based evaluation in a market-oriented school system. In M. Dahlstedt & A. Fejes (Eds.), *Neoliberalism and market forces in education: Lessons from Sweden* (1st ed.). Routledge. https://doi.org/10.4324/9780429470530

Yoshikawa, H., et al. (2015). Experimental impacts of a teacher professional development program Chile on preschool classroom quality and child outcomes. Developmental Psychology, 51, 309–322.

Zentall, S. S., & Javorsky, J. (2007). *Professional development for teachers of students with ADHD and characteristics of ADHD. Behavioral Disorders, 32*(2), 78–93. https://doi.org/10.1177/019874290703200202

Supplementary material

Supplementary material section A

Table A1

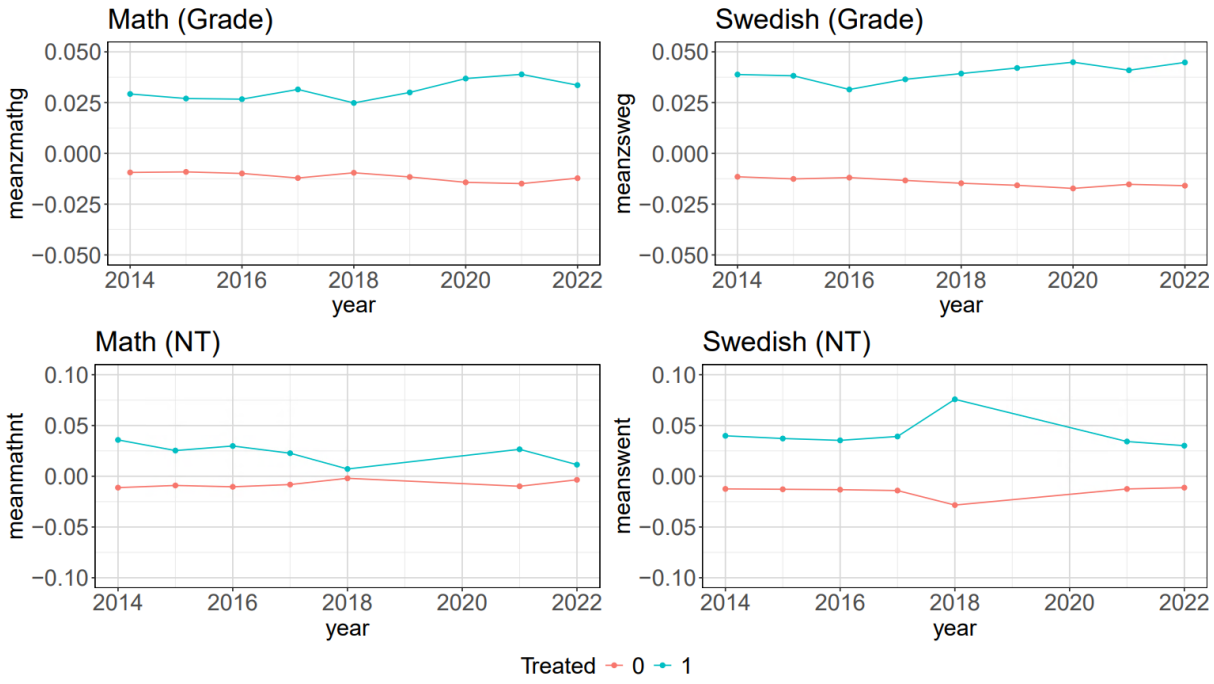
Pretend test

	Chi square	*p*-value
Grades		
Math	26.29	0.557
Swedish	24.22	0.667
National test scores		
Math	35.19	0.0370
Swedish	18.79	0.6578

Note. The table reports Chi square statistics and corresponding p-values from pre-treatment tests of differential trends between treatment and control schools. Non-significant results indicate no evidence of pre-treatment differences in trends, consistent with the parallel trends assumption.

Figure A1

Outcome variables over time



Note. No national tests were conducted in 2019 or 2020 due to the coronavirus pandemic.

Supplementary material section B

Figure B1

Students with individualized education programs (IEPs)

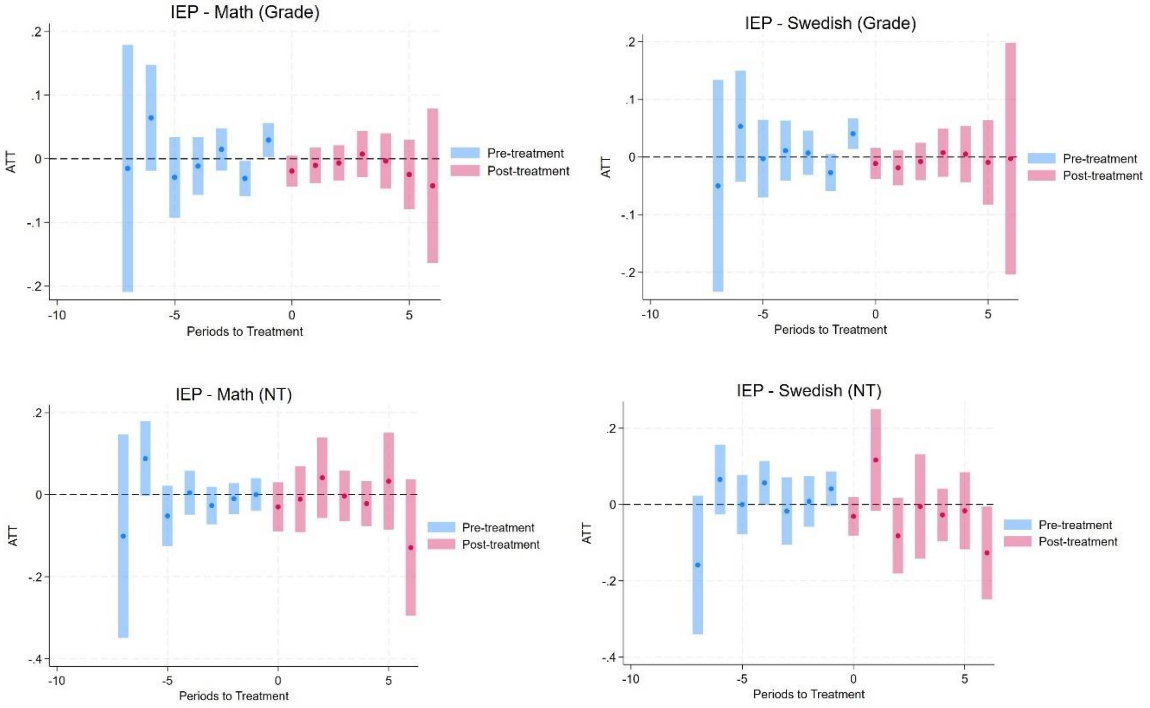


Figure B2

Boys

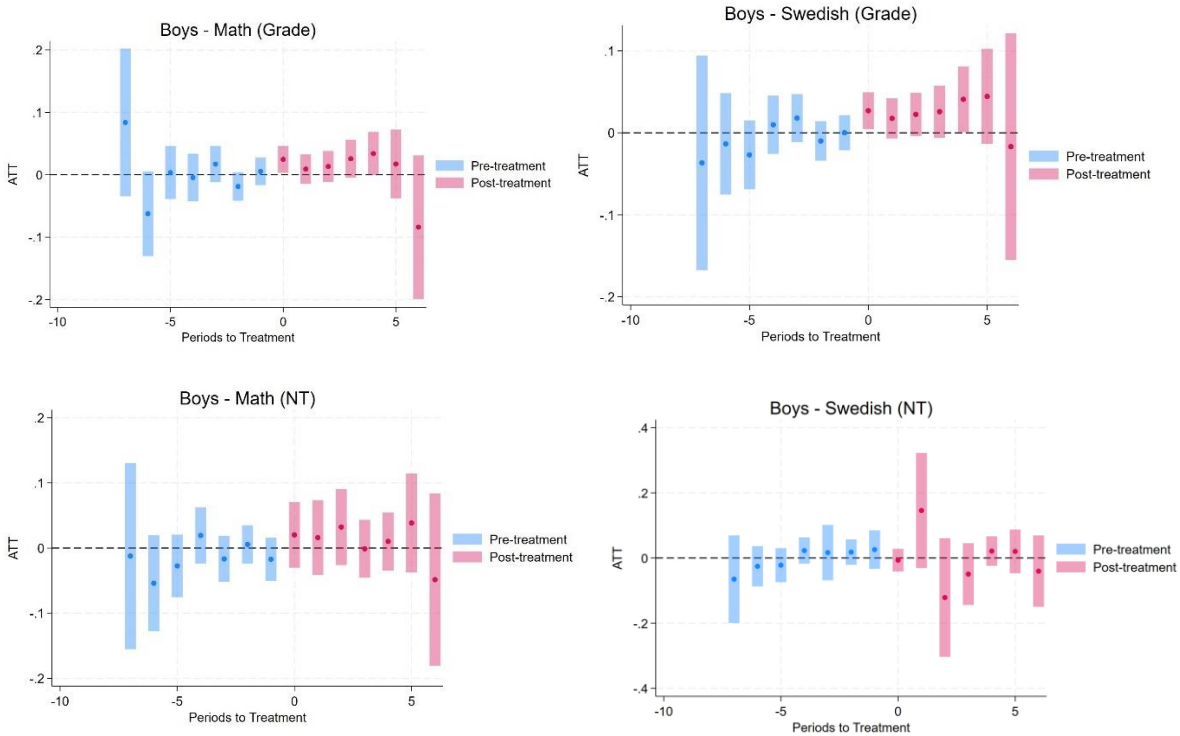


Figure B3

Girls

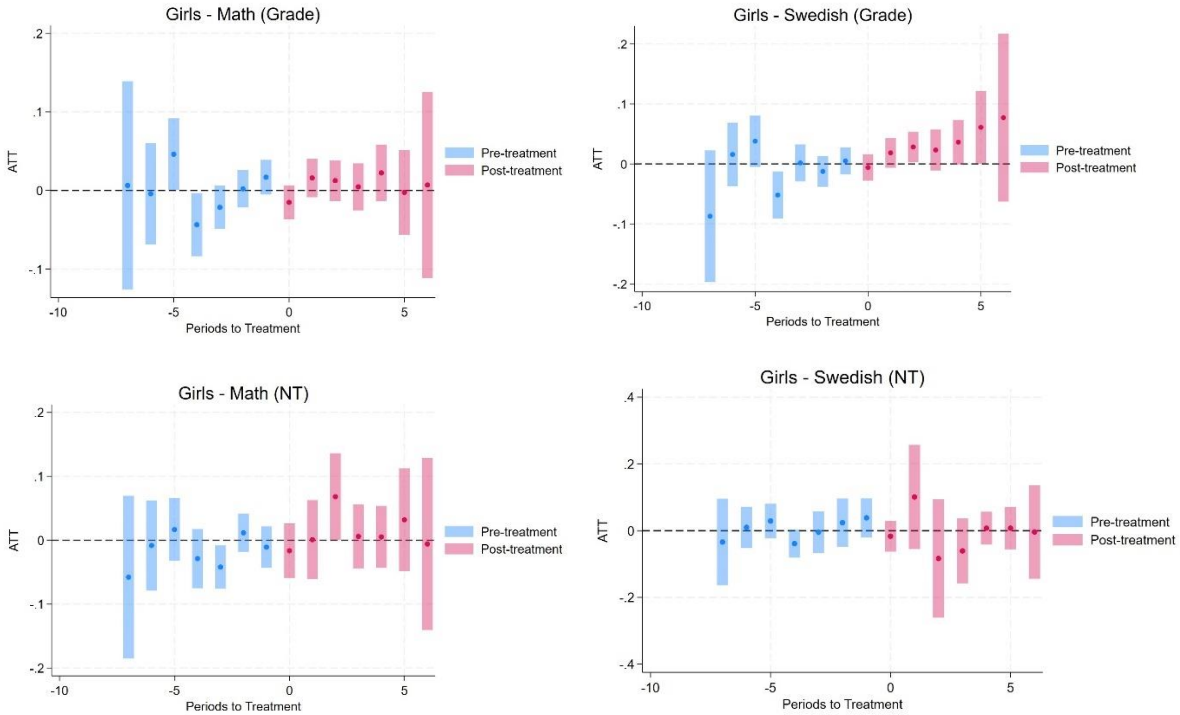


Figure B4

Migrant background

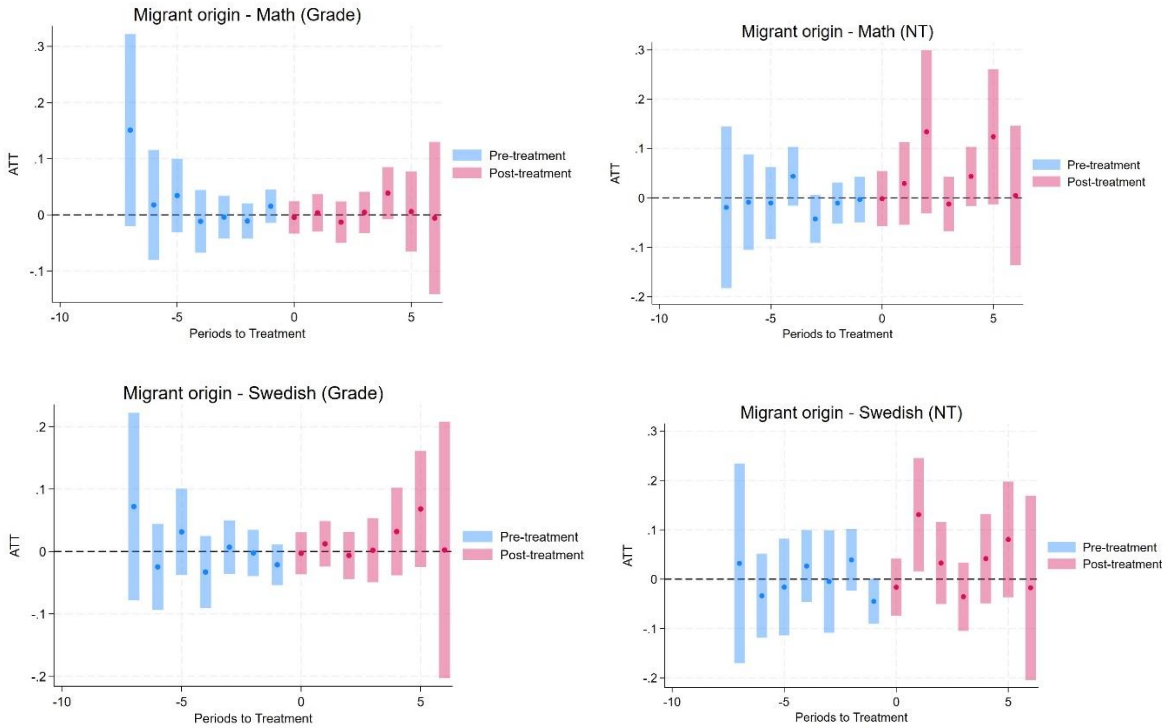


Figure B5

Childhood immigrants

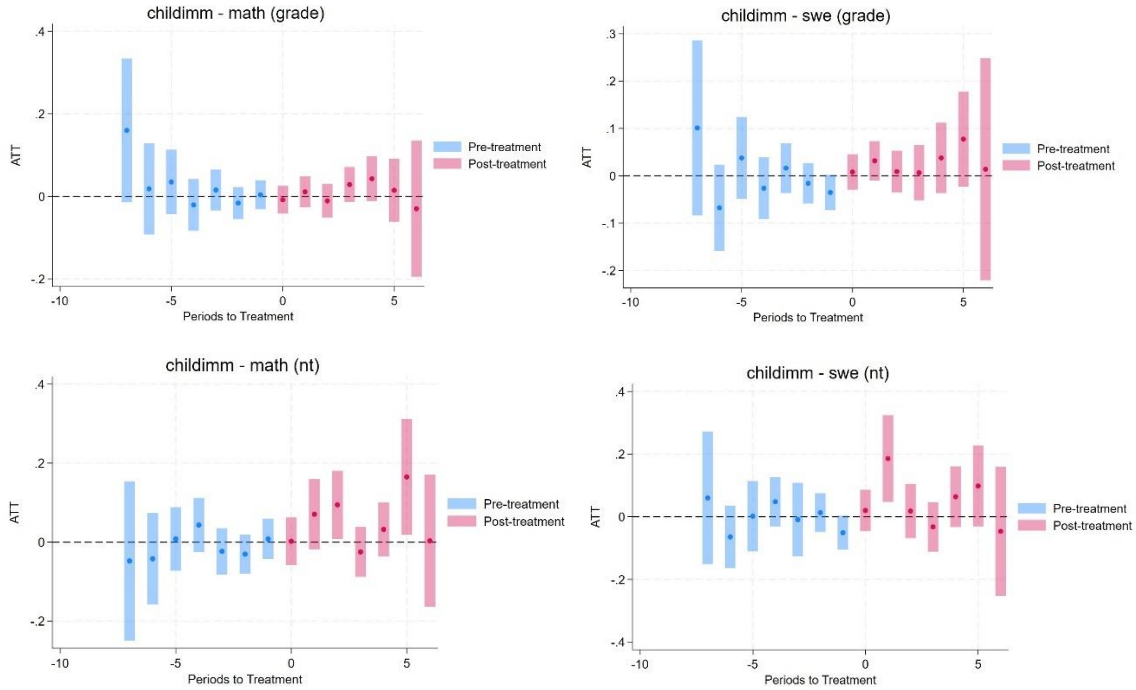
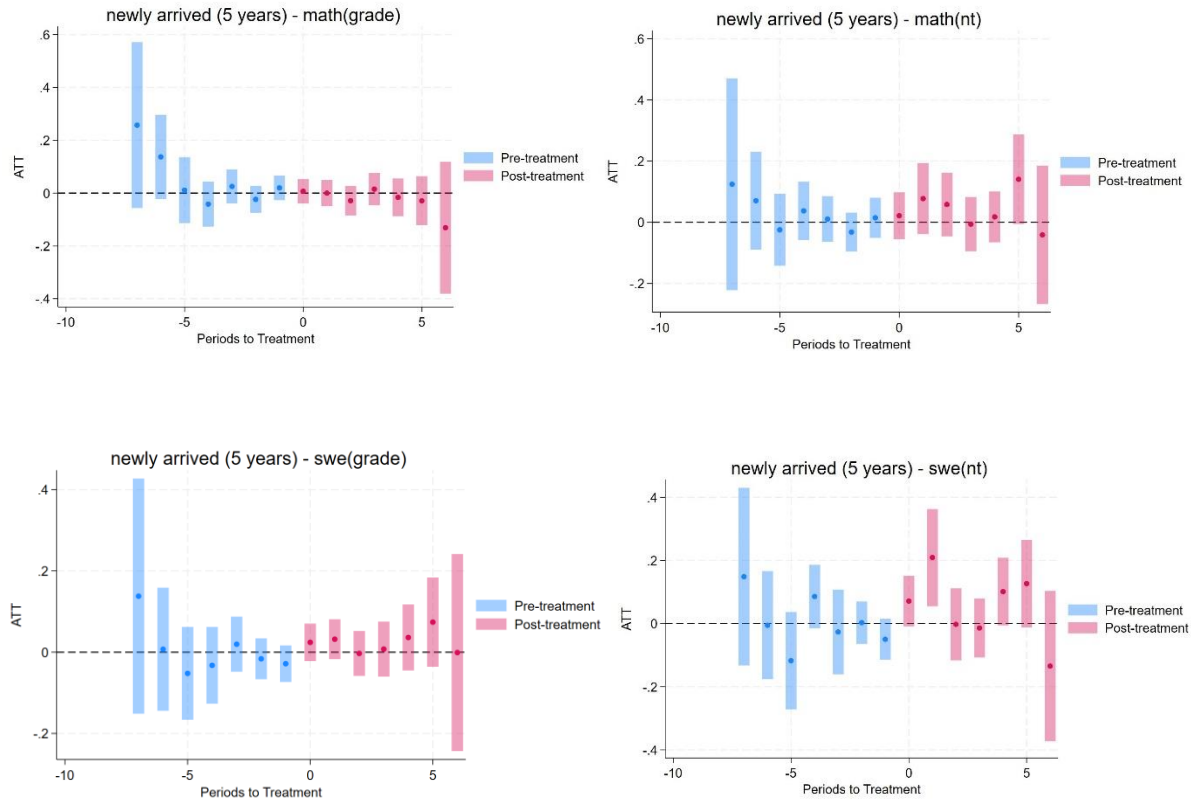


Figure B6

Arrived within 5 years



Supplementary material section C

Figure C1

Bottom achievement quartile

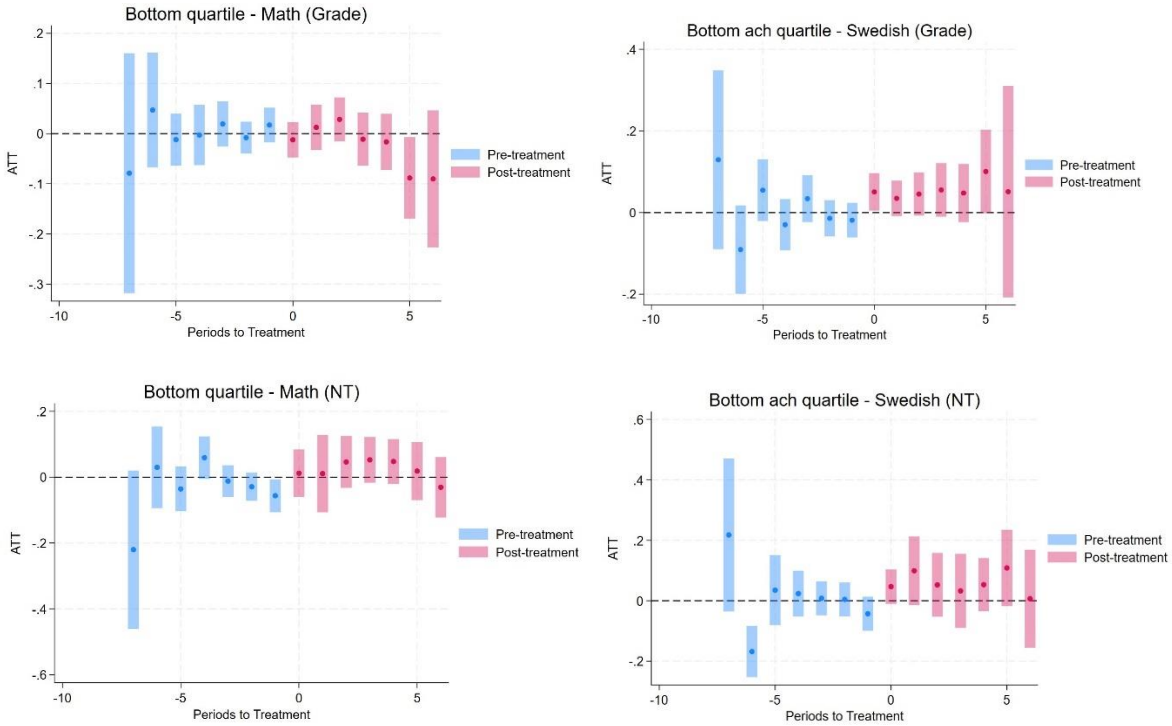


Figure C2

Top achievement quartile

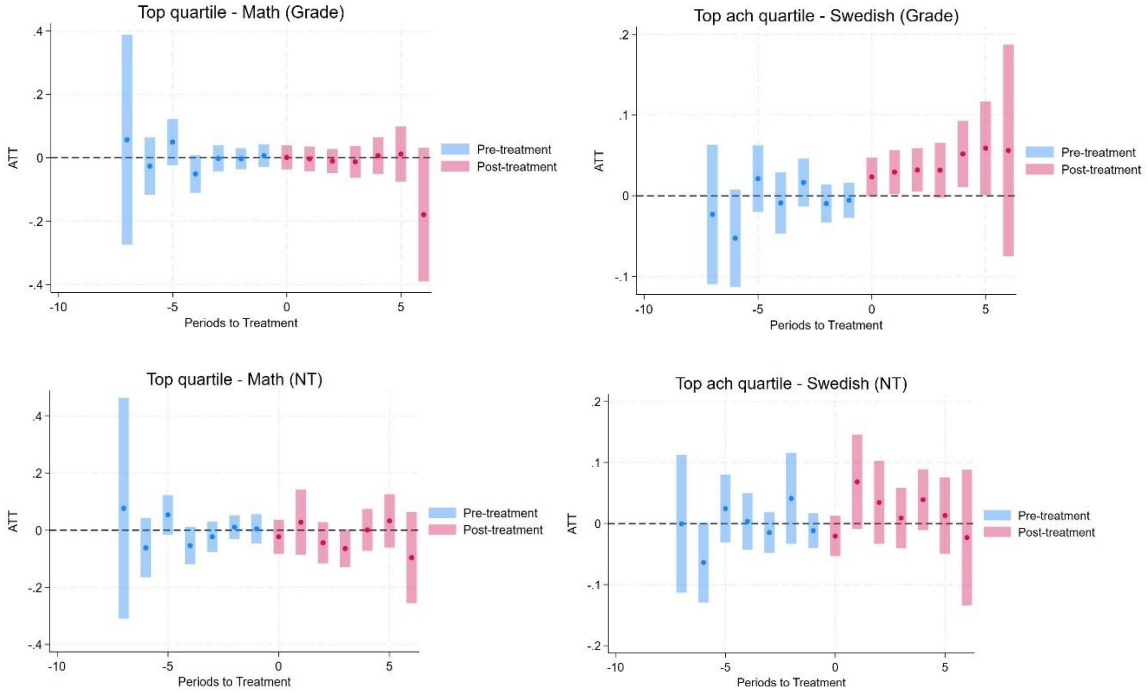


Figure C3

Independent schools

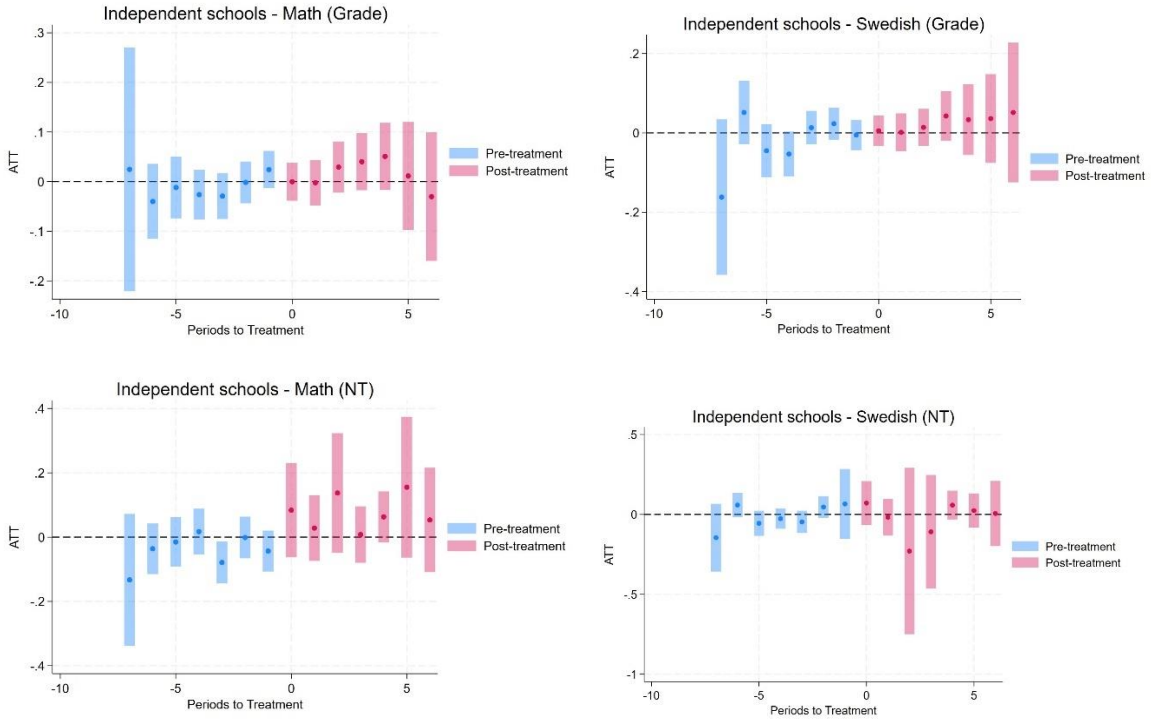
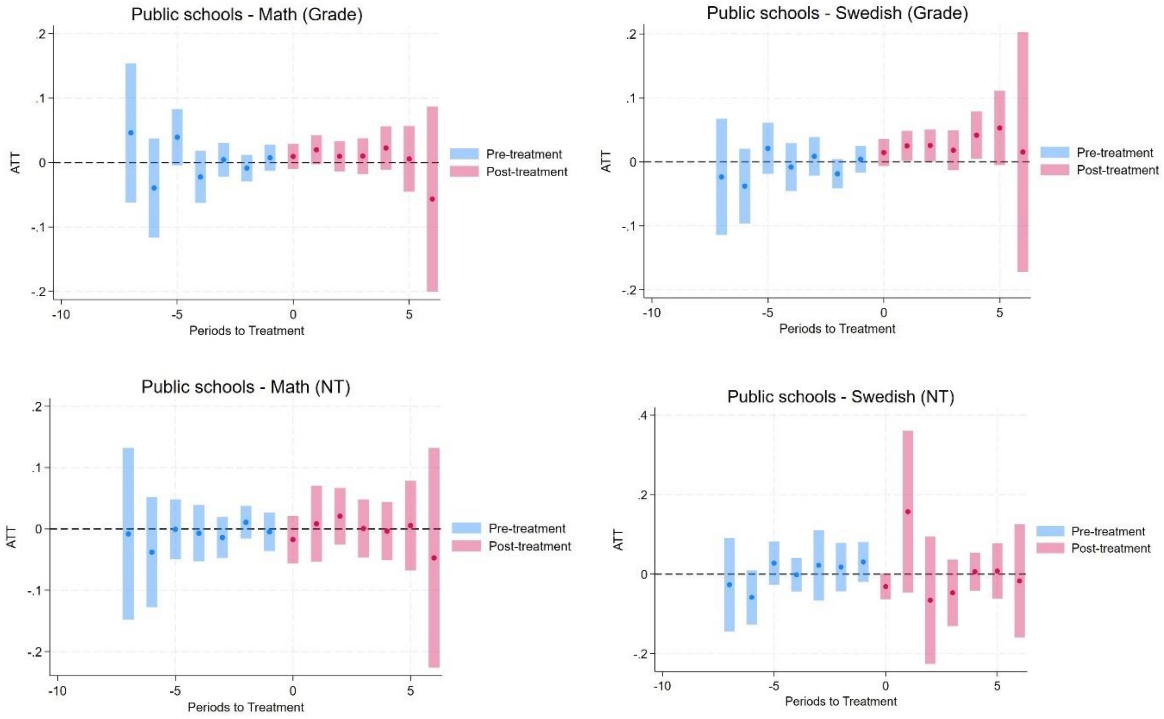


Figure C4

Public schools



Supplementary material section D

Figure D1

Alternative treatment specification - estimated event time effects

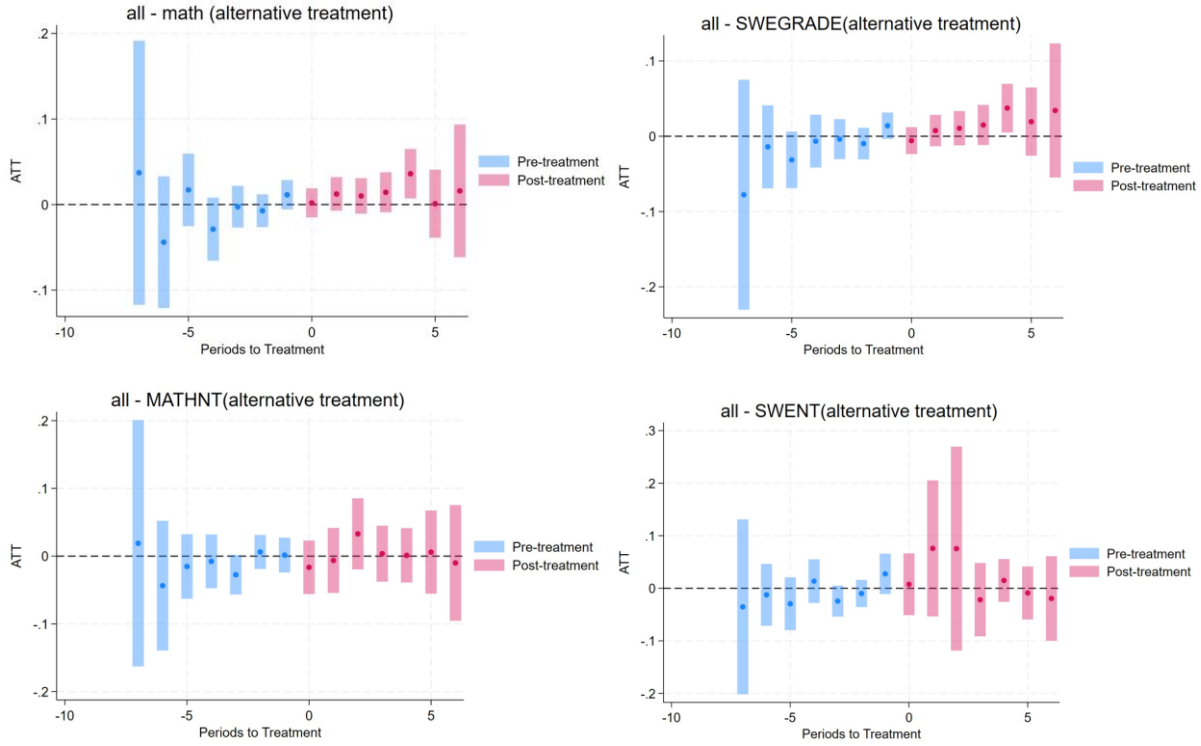


Table D1

Estimated effects of SFL for all students (alternative treatment specification)

	Coefficient	Standard Error	z	p-value
Grades				
Math	0.0123	0.008	1.39	0.163
Swedish	0.0119	0.010	1.17	0.241
National test scores				
Math	-0.0001	0.015	-0.01	0.992
Swedish	0.0169	0.021	0.79	0.427

Note. The table shows the estimated effects of SFL on pooled Grade 6 and grade 9 math and Swedish grades and national test scores across all students. Scores have been standardized by outcome measure and year.

Table D2

Estimated effects of SFL for student and school subgroups based on alternative treatment definition (national test scores only)

	Coefficient	Standard Error	z	p-value
IEP				
Math	-0.001	0.028	-0.04	0.968
Swedish	0.006	0.032	0.21	0.837
Arrived within 5 years				
Math	0.025	0.031	0.81	0.418
Swedish	0.061	0.039	1.57	0.116
Independent schools				
Math	0.029	0.032	0.88	0.379
Swedish	0.042	0.033	1.26	0.207
Low certified teachers				
Math	0.045	0.042	1.13	0.260
Swedish	0.019	0.040	0.49	0.628

Note. The table shows the estimated effects (ATTs) of SFL on grade 6 and grade 9 math and Swedish national test scores. Outcomes have been standardized by grade and yearly cohort. The change in effect column refers to the change in effect as compared to the main treatment definition (conditioned on a coach being assigned to the school). The sensitivity checks are only run for subgroups where statistically significant effects were found in the model specification including main treatment definition.

Supplementary material section E

Table E1

Estimated effects of SFL for students in small schools

	Coefficient	Standard Error	z	p-value
Grades				
Math	0.0252	0.0222	1.13	0.258
Swedish	0.0113	0.0272	0.42	0.767
National test scores				
Math	0.023	0.0455	0.52	0.600
Swedish	-0.018	0.0369	-0.49	0.625

Note. Schools are classified as small if they fall within the bottom quartile of student enrollment (a pre-treatment 2-year average). Coefficients are ATTs.

Supplementary material section F

Table F1

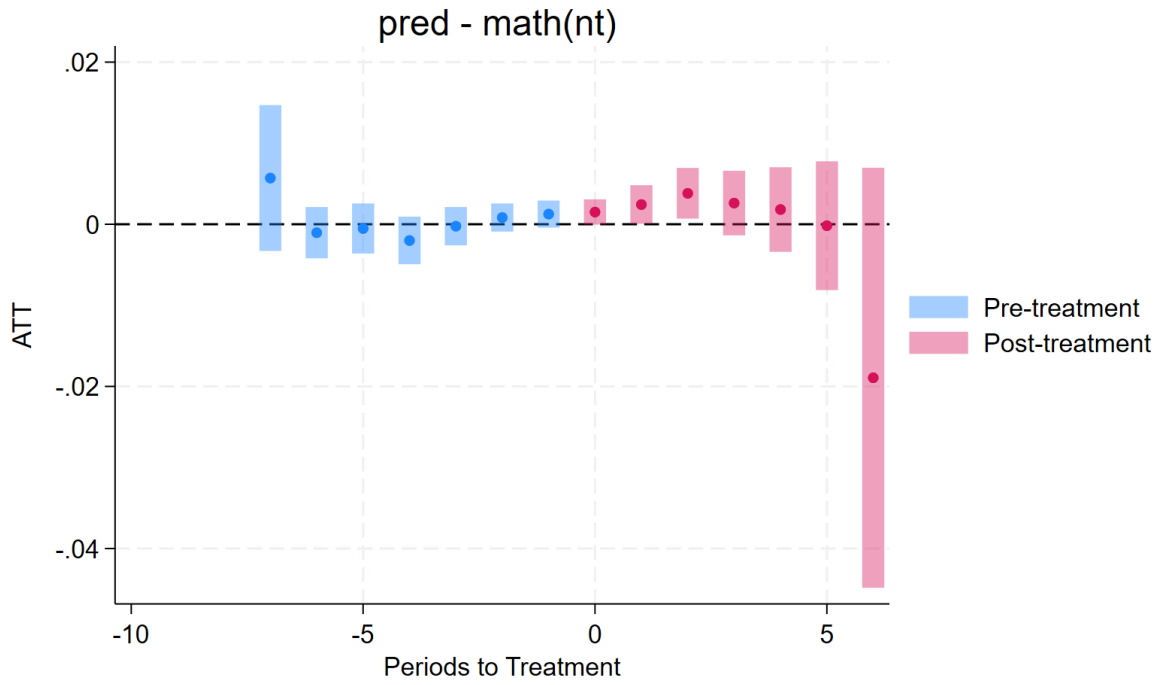
Estimated effects of SFL on predicted mathematics national test scores

	Coefficient	Standard Error	z	p-value
Outcome				
Predicted math test score	0.002	0.0014	1.46	0.145

Note. Variables in the linear prediction models include gender, migration background, highest level of parental education, and school fixed effects. $R^2 = 0.16$.

Figure F1

Event study estimates of SFL on predicted test scores



Supplementary material section G

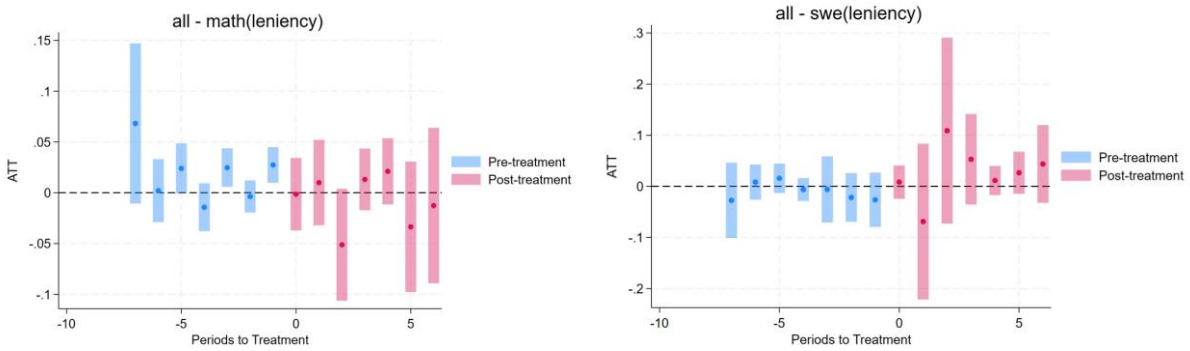
Table G1

Grading leniency

	Coefficient	Standard Error	z	p-value
Math	-0.0019	0.0123	-0.15	0.877
Swedish	0.0197	0.0209	0.94	0.346

Note. Grading leniency is operationalized as the difference between cohort-year standardized grades and national test scores.

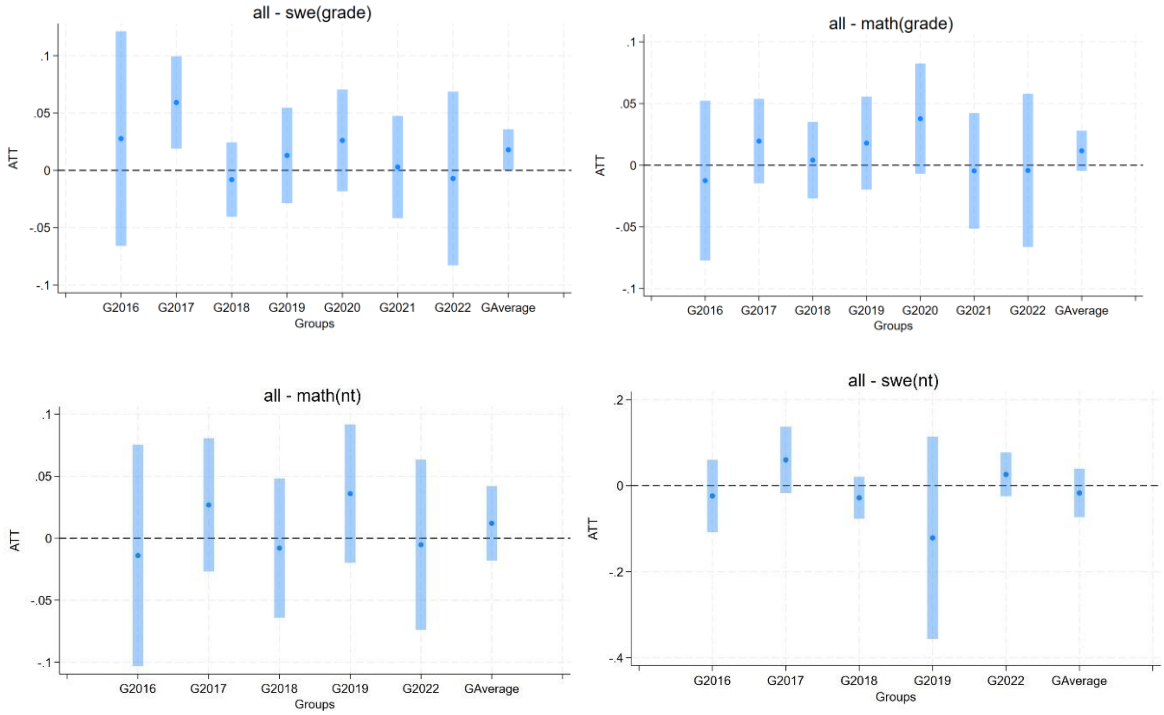
Figure G1



Supplementary material section H

Figure H1

Cohort-specific effects (all students)



Supplementary material section I

Figure I1

All students

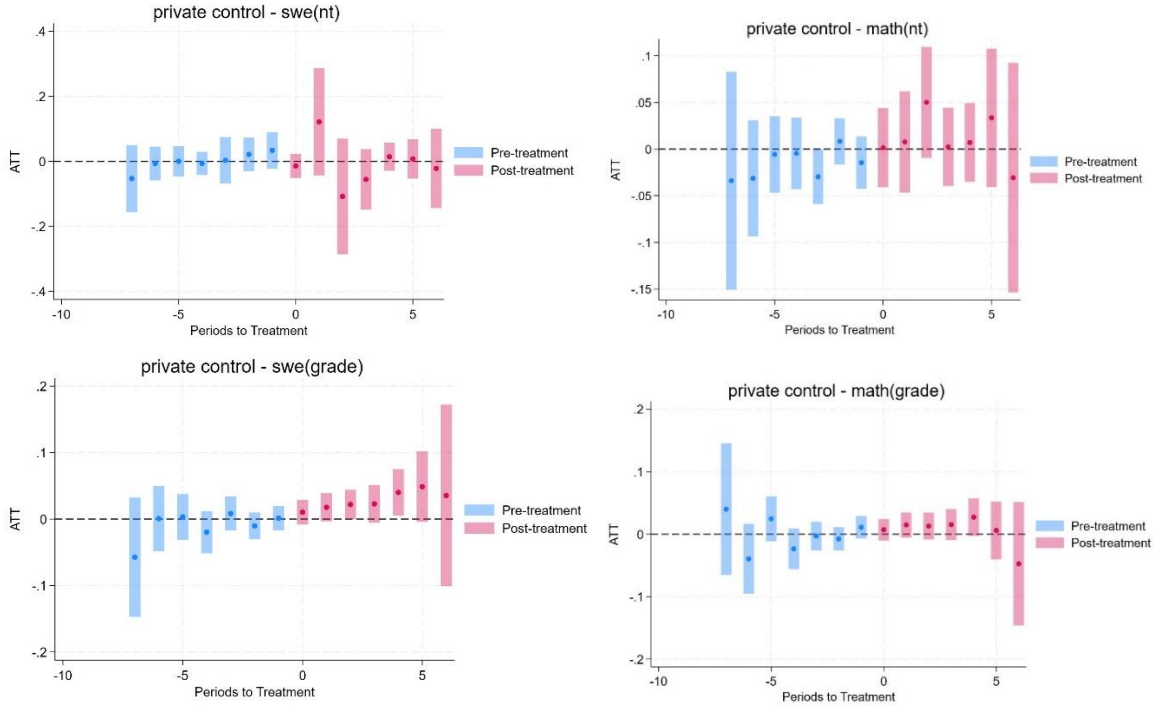


Figure I2

Students with IEPs

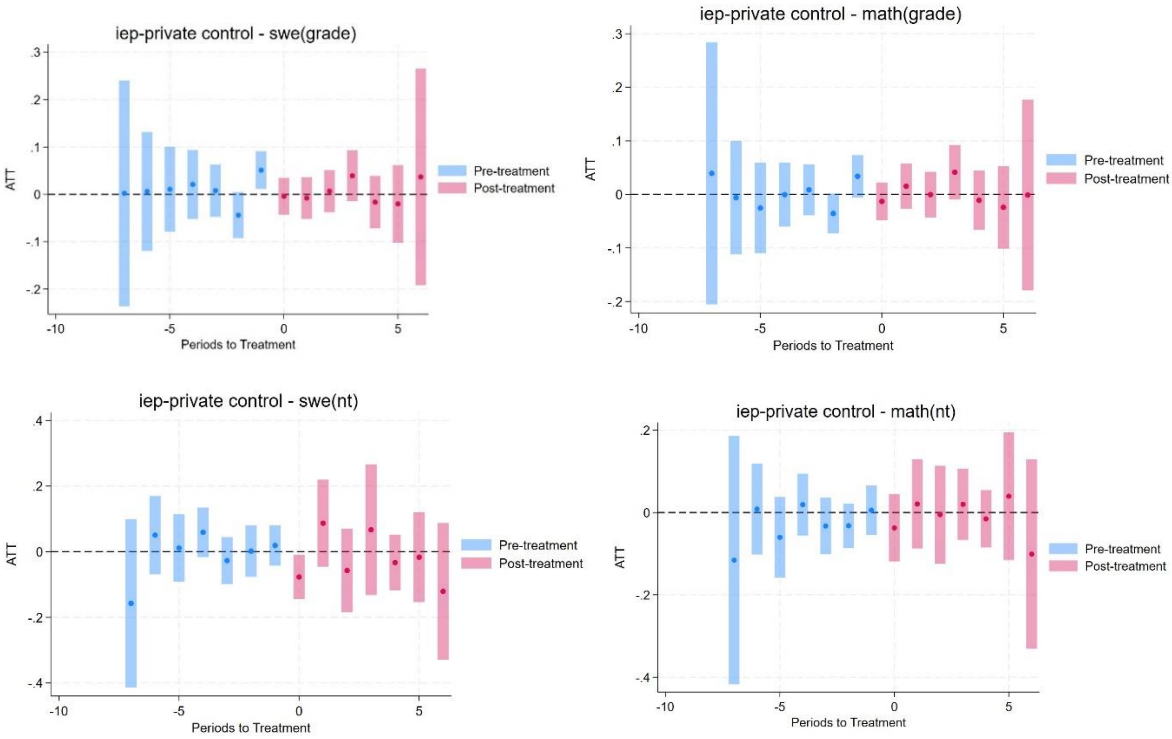


Figure I3

Students arrived within 5 years

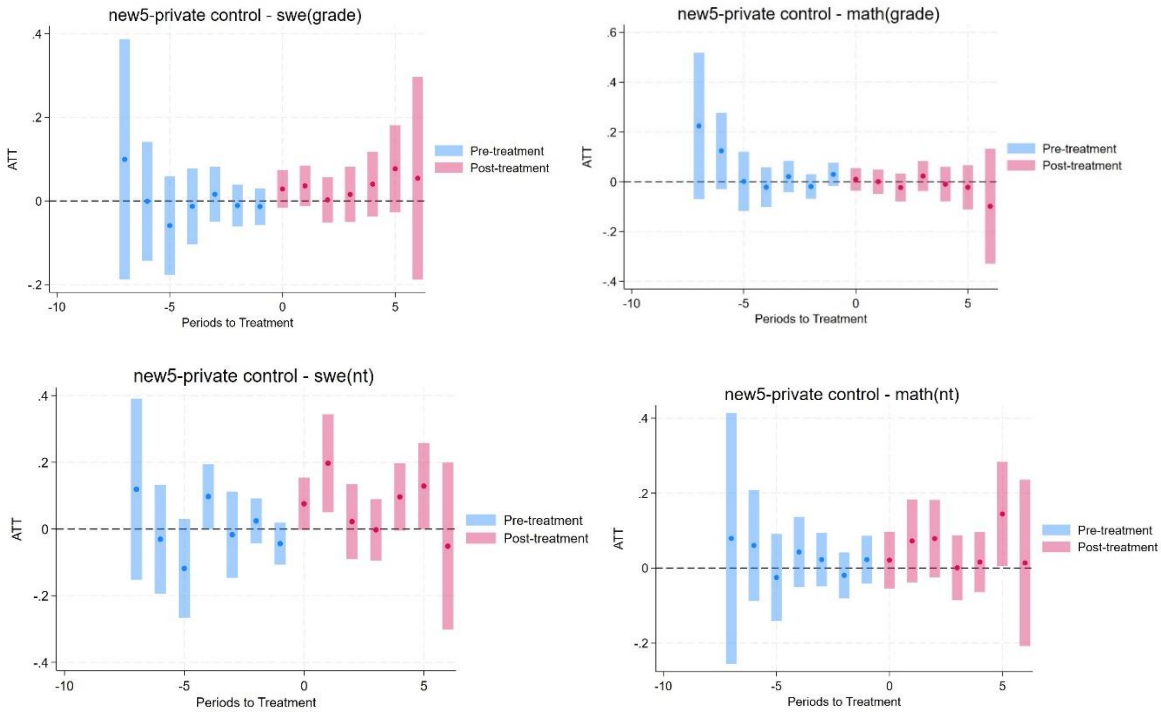


Figure I4

Childhood immigrants

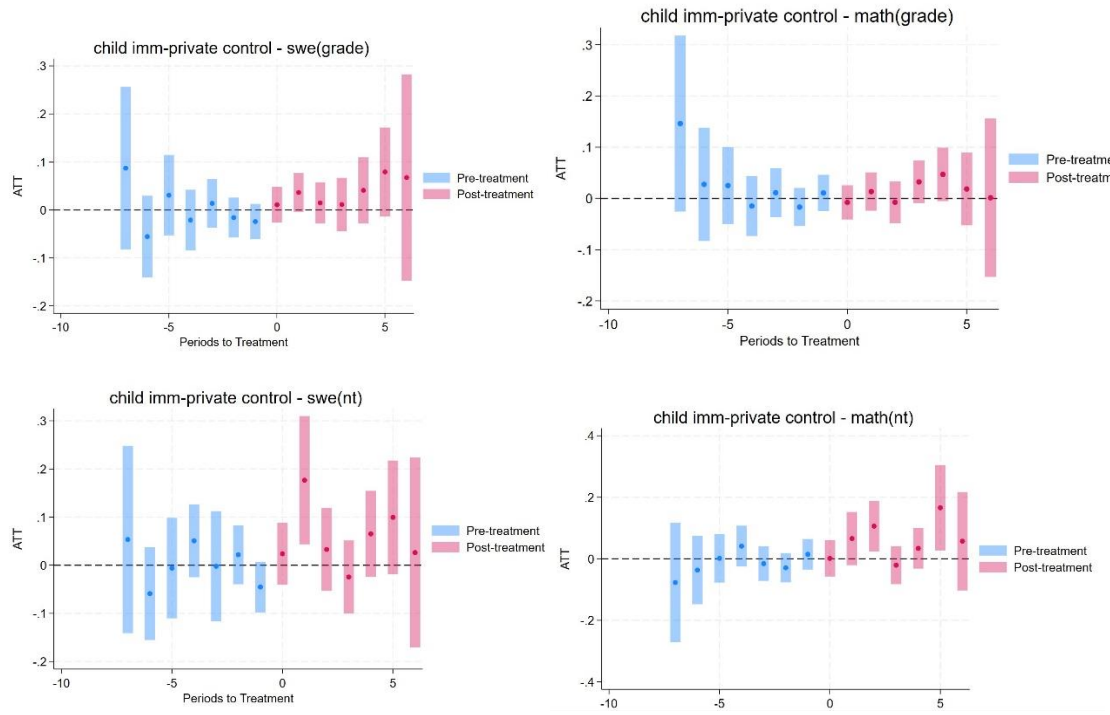


Figure I5

Migration background