Do funds for more teachers improve student outcomes?[*]

Nicolai T. Borgen[†], Lars J. Kirkebøen[‡], Andreas Kotsadam[§] and Oddbjørn Raaum[¶]

November 28, 2024

Abstract

We investigate the effects of a large-scale Norwegian intervention that provided extra teachers to 166 lower secondary schools with relatively high student-teacher ratios and low average grades. We exploit these two margins using a regression discontinuity setup and find that the intervention reduced the student-teacher ratio by around 10%, without crowding out other school resources. The extra funding did not improve test scores and medium-term academic outcomes, and we can reject even small positive effects. We do find that more teachers improved the school environment, including self-reported student well-being, but with the largest impact on aspects of the school environment most weakly associated with better academic outcomes.

Keywords: Student-teacher-ratio, class size, test scores, non-cognitive skills, RDD
JEL codes: J24, I2

[*] We thank seminar participants at several universities and conferences, Martin Andresen, Colin Green, Astrid Sandsør and Diane Schanzenbach for useful comments and suggestions. This research is funded by the Norwegian Directorate of Education and Training, the European Research Council (grant # 818425), and the Research Council of Norway (grant # 300917), and is partially supported by the Research Council of Norway through its Centres of Excellence scheme (grant # 331640).

[†] University of Oslo; e-mail: n.t.borgen@isp.uio.no

[‡] Statistics Norway, Oslo; e-mail: lars.johannessen.kirkeboen@ssb.no

[§] Ragnar Frisch Centre for Economic Research, Oslo; e-mail: andreas.kotsadam@frisch.uio.no

[¶] Ragnar Frisch Centre for Economic Research, Oslo; e-mail: oddbjorn.raaum@frisch.uio.no

1 Introduction

There is a long-standing debate as to whether reducing class size in rich countries improves student outcomes or not (e.g., Angrist and Lavy, 1999; Angrist et al., 2019; Browning and Heinesen, 2007; Fredriksson et al., 2013; Leuven and Løkken, 2020). However, the effects of reductions of class size, or the student-teacher ratio, are challenging to identify because class size is potentially endogenous with respect to student characteristics (Lazear, 2001), and more teachers per student may influence decisions by parents, schools, and education authorities regarding other inputs (Todd and Wolpin, 2003). Since large RCTs such as Project STAR (Krueger, 1999) are rare, the literature is dominated by natural experiments such as rule-induced class size, close elections, or other exogenous sources of differential funding of school districts. Not only have these studies provided mixed findings (see below), but the variation in student-teacher ratios used for identification is often far from what is relevant for large-scale policy.

We provide direct evidence of the short- and medium-term effects of additional school funding, earmarked for hiring more teachers. We investigate the effects of a large Norwegian intervention that provided four-year funding of 600 extra teachers to 166 lower secondary schools with a student-teacher ratio higher than 20 and a student grade point average below the national mean. By exploiting these two margins in a regression discontinuity (RD) setup, we find that the intervention reduced the student-teacher ratio by around 10% (from 22 to 20), without crowding out other school resources, reducing teacher qualifications, or diminishing parental support.

The funding had no impact on academic outcomes. More teachers did not raise test scores after one year, externally graded exam scores after three years, or later completion of upper secondary education. Furthermore, we can reject even small positive effects. For example, the upper bound of the funding effect on test scores is just 3% of a standard deviation, which is considerably smaller than what we should expect based on e.g. Project STAR (Krueger, 1999). However, we do find that the increased number of teachers improved the school environment, including student well-being.

The intervention provides an excellent opportunity to identify the effects of additional funding for more teachers. First, the policy substantially increased the number of teachers and was implemented on a large scale. The policy increased the annual costs per student by USD 1,400 over four years in the targeted schools, from an already high level of spending. Second, the policy was implemented in a way that permits credible identification. The two sharp margins of the eligibility criteria enable us to use an RD framework, which can credibly address concerns about omitted variable bias. Third, the policy effects can be investigated using rich register data, allowing us to study both short- and medium-term academic outcomes, with minimal attrition and non-response problems. The total size of the intervention (166 schools over four years, with a total of about 15,000 students per cohort) gives us sufficient precision to rule out even small effects. The data also enables us to examine compensatory adjustments by parents or schools induced by the policy, and to check whether the policy impacted recruitment of qualified teachers. Moreover, by supplementing register data with survey data, we report effects on students' perception of their school environment, including measures such as support from teachers, well-being, and bullying.

By leveraging a policy that allows us to credibly identify the effects of marginal educational investments in lower secondary teachers in Norway, our results are primarily relevant for policymakers in similar contexts, where expenditures are already high. Although Norway is among the high-spending countries, total spending is not very different from many other rich countries. Moreover, the average class size in Norwegian lower secondary schools is about 24.5 (Falch et al., 2017; Leuven and Løkken, 2020), which is similar to the US (25) and larger than the OECD average of 23 (OECD, 2021). Thus, our results are relevant for policymakers' decision on marginal investments in lower secondary school teachers in many OECD countries.

Our main contribution is the identification of a policy-relevant parameter using marginal educational investments. While natural experiments credibly identify class size effects, they often study variation in student-teacher ratios at margins less relevant for marginal educational investments. This is so even when they use policy induced variation, as in Jepsen and Rivkin (2009) who find that reducing class sizes in California from 30 to 20 students raised mathematics and reading skills. In addition, class size reductions triggered by maximum class size rules could be subject to various input substitutions, such as providing fewer teacher hours or less qualified teachers in smaller classes. From a policy perspective, such endogenous inputs are an intrinsic part of the policy effect (Todd and Wolpin, 2003) and should clearly not be held constant (Leuven et al., 2008). However, the specific input substitutions (or absence thereof) may be closely related to the specific class size margins that are studied. Moreover, the effects of a given class size reduction may differ depending on whether the initial class size is 40, 30, or 20. Thus, class size effects triggered by maximum class size rules may differ from the effects of marginal changes in the student-teacher ratio.

This paper adds to the inconclusive literature on class size effects by providing direct evidence on effects of additional funds to hire more teachers. In a landmark study, Krueger (1999) investigated the effects of Project STAR, which randomly assigned students to smaller classes from kindergarten through third grade, and found a large improvement in performance due to class size reductions. Later work investigated the longer-term effects of the STAR experiment and found positive effects on college attendance (Chetty et al., 2011). Conversely, Hoxby (2000) exploits "as if random" variation across cohorts in Connecticut schools and rules out even small effects of class size on performance in math, reading, and writing in fourth and sixth grade. Woessmann and West (2006) also use variation across grades within schools to identify class size effects and only find effects in 2 of 11 countries using TIMSS data.

Most studies investigating the effects of class size use rule-induced reductions, where classes need to split if they reach a certain threshold. A classic example of this method is Angrist and Lavy (1999), which has been interpreted as finding positive effects due to smaller classes in Israel, although the results were actually mixed. They found consistent effects for fifth graders, mixed results for fourth graders, and zero effects for a sample of third graders. Angrist et al. (2019) find no effects using later cohorts in a follow-up study. Results from other contexts are also mixed. Positive effects due to class size reductions have been found in Sweden (Fredriksson et al., 2013), Denmark (Browning and Heinesen, 2007), and Bolivia (Urquiola, 2006). In contrast, Argaw and Puhani (2018) find no effect of class size in elementary school on choosing a more academic track in Germany, and Angrist et al. (2017) find no effects on gains in learning in Italy.

Likewise, previous Norwegian studies using rule-induced reductions found mixed results. An early study by Bonesrønning (2003) found some weak evidence that larger classes in lower secondary school lead to less favorable short-run outcomes, but no subsequent Norwegian study has found the same. Leuven et al. (2008) find no effect on short-run test scores in lower secondary schools. Leuven and Løkken (2020) reject small effects on long-run outcomes from class-size reductions in primary and lower secondary schools, and Falch et al. (2017) find no long-run effects from size reductions in lower secondary schools. Thus, in Norway, the effects of class size reductions triggered by maximum class size rules and marginal educational investments appear to be the same.

We also add to the broader literature on the effects of school resources on educational outcomes.[1] Recent US studies using school financing reforms or changes in the components of the school financing formula show positive effects on test scores and longer-run educational attainment (Jackson, 2020; Jackson and Mackevicius, 2021; Deming, 2022). From other rich countries, there are fewer studies on the effects of non-targeted resources with credible research designs. One exception is Hægeland et al. (2012), which find that adding resources stemming from hydro-power revenues to schooling in Norway positively affected learning outcomes. Another exception is Gibbons et al. (2018), who use boundary discontinuities in England and find that higher spending leads to better educational attainment. Our estimates add to the small number of studies credibly investigating the effects of spending in countries other than the United States.

Finally, our study contributes to the burgeoning literature on the marginal effects of resources on non-test score outcomes and the school environment. Test scores do not capture all that students learn in school. This is most clearly evinced by the fact that successful interventions often have even larger effects on longer-run outcomes, such as the probability of attaining higher education, than they do on immediate test scores (e.g., Jackson 2018; Jackson et al. 2020). Therefore, an increasing number of recent studies aim to identify effects on non-test-score outcomes, which could be caused by a wide variety of skills, including what are often labeled non-cognitive skills. Schools can affect non-cognitive skills, which often have long-term impacts on educational and labor market outcomes (Chetty et al., 2011; Cornelissen and Dustmann, 2019; Heckman et al., 2013; Fredriksson et al., 2013). US evidence suggests that schools that improve ninth-grade socioemotional development also increase upper secondary completion and college enrollment. Jackson (2018) finds that teachers affect non-cognitive skills and that these teacher-induced improvements in non-cognitive skills are better predictors of longer run academic outcomes than teacher-induced changes in test scores. We are not aware of any study testing the effects of reduced class size on school environment outcomes, but Chingos (2012) finds that smaller classes in California led to reduced absenteeism and Dee and West (2011) find that smaller classes make students less afraid to ask questions and more likely to look forward to a subject. Neither of the two studies find effects on test scores.

The remainder of the paper proceeds as follows: We describe the institutional context and data in Section 2 and present the empirical approach in Section 3. We present the estimated effects on student teacher ratios in Section 4, academic outcomes in Section 5, and school environment in Section 6. Section 7 discusses the magnitudes of the effects and different explanations for the zero effects on academic outcomes, while Section 8 concludes.

[1] See also Baron (2022); Abott et al. (2020); Brunner et al. (2020, 2022) for recent studies that investigate what type of spending matters or in which context the effects are largest.

2 Context, intervention, and data

Up to tenth grade (age 16), municipalities administer the public schools in Norway. The municipalities acquire revenue from income tax, user charges on services, and transfers from the central government. They spend money on schools as well as on child care, health care, and other services like water supply and renovation. Schools are free of charge and compulsory through grade 10, and the private share of lower secondary schools is small (about 3-4%[2]).

Between-school differences are relatively modest in Norway, including differences in resources. The allocation of input in terms of teaching personnel is highly compensatory. To reduce achievement gaps, school administrations in some large municipalities allocate more resources to schools with disadvantaged student populations. As a result, students from low-income families with less-educated parents typically attend smaller classes (Leuven et al., 2008). In our data we can illustrate this compensatory policy by means of entry test scores in grade 8 (lower secondary school). A difference in school level entry test scores of 0.1 standard deviation is associated with 0.35 fewer students per teacher in grade 8-10, compared to a mean of 15.9 (panel A of Appendix Table A.1).

In the fall of 2012, the Norwegian Parliament decided to reinforce this policy by providing a four-year extra funding for about 600 more teachers in grades 8 to 10 per year to municipal school administrations (NOK 1.5 billion, or USD 258 million using the 2012 exchange rate), starting with the 2013/2014 cohort. In line with the tradition for compensatory resource allocation, the extra resources were channeled to 166 lower secondary schools with an average student-teacher ratio for regular instruction[3] above 20 *and* average grades at the end of grade 10 below the national mean in the previous school year (2011/2012). Thus, the schools could not manipulate their distance to the cutoffs. For schools that met both conditions, the number of extra teachers varied according to school size. Schools received funds for one, two, three, four, or five additional teachers depending on whether the number of students was 0-99, 100-199, 200-299, 300-399, or more than 400. The schools with extra funding were located in 98 different municipalities, and all counties in Norway were represented.

The purpose of the intervention was to enable schools to better tailor the teaching to the individual students and thereby raise basic skills, improve the learning environment, and reduce special needs education. The extra resources were intended to increase the number of qualified teachers in regular teaching and were not to be used for special needs education. Apart from these requirements, schools were free to organize the regular teaching as they wanted (two teachers in regular classes, divide classes into smaller groups, etc.), and they could use the funding in different ways across subjects, grades, and time. A survey administered to the principals at all treatment schools (with a 65% response rate) provides some information on how schools used the extra funds (Kirkebøen et al., 2017). It should be noted that since the resources could have been used differently across subjects, grades, and time, principals had the option to report multiple measures, and the survey only offers broad indications of the extent to which individual students experienced

[2] Statistics Norway, table 05232 (https://www.ssb.no/statbank/table/05232).

[3] The student-teacher ratio is measured as the number of regular-instruction student hours divided by the number of regular-instruction teacher hours. Thus, it disregards student and teacher hours spent on for example special needs teaching and special services for Norwegian language learners, and measures of the average group size in regular instruction settings.

each measure. About three out of four principals (77%) report that the extra funds were used to have two teachers available for the class, while 66% report that the funds were used to divide the class into smaller groups. About 40% reported using the resources to provide small-group instruction, and less than 15% used the resources for extra instruction for individual students. Very few principals report that the extra resources were always used to divide groups by students' skill levels. However, nearly 60% report that groups were sometimes divided in this way (with 40% reporting that it was never divided according to skills level). The principals report that the resources were primarily used in math (95%), Norwegian (90%), and English (78%). Finally, 70% report that resources were shared across all lower secondary grades. Given our precise null findings documented below, it is also interesting to note that 95% of principals believed that the funding improved students' learning outcomes, suggesting that the perceived effectiveness of interventions may differ considerably from the evidence based on a credible treatment effect analysis.

The starting point for our data set is the population of eligible lower secondary schools. The extra funding was allocated on the basis of 2011/2012 regular student-teacher ratios and grade point averages (GPAs) based on teacher-assessed subject-specific performance and externally-graded exams scores by the end of grade 10. Teacher grades dominate the GPA, but the written and oral exams also count (weight of about 0.1). Student-teacher ratios and GPAs are well defined for 859 out of 1 089 public lower secondary schools, containing 97.5% of the students. The remaining 230 schools, which are small, many of which cater for special needs students, are excluded.

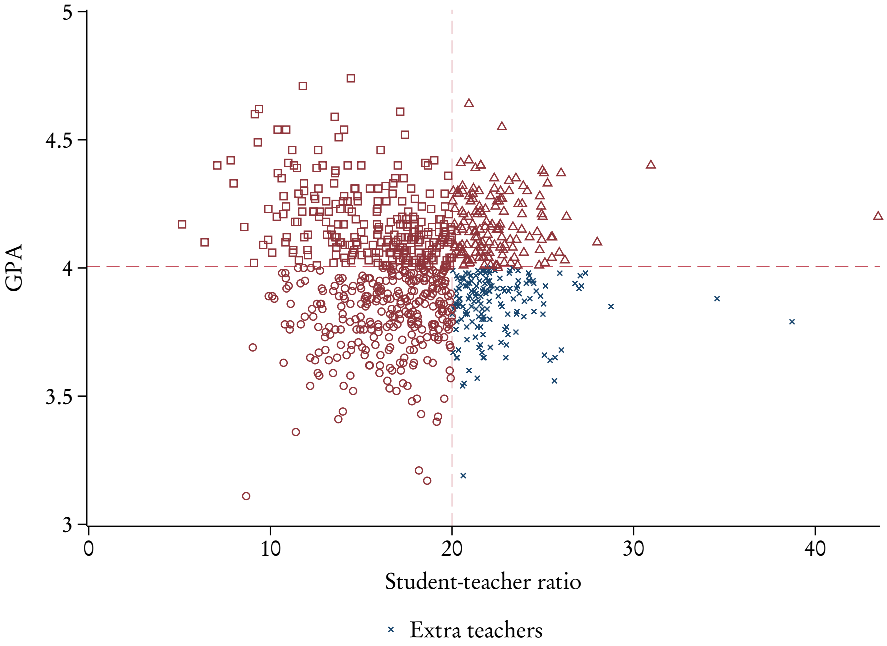
Figure 1 shows how schools are distributed along the student-teacher ratio (x-axis) and the GPA (y-axis) dimensions in the pre-reform year (2011/2012). The treated schools, marked with blue x, have below-average GPAs and group size above the average. The two sharp margins enable us to use a regression discontinuity framework, as discussed below.[4]

We merge our school treatment status data set with four different data sources that together allow us to study a variety of outcomes: (i) the compulsory school register ("GSI"), with school-level data enabling us to study resource use in schools, including student-teacher ratios, the share of qualified teachers, and how teacher resources are used (e.g., on regular teaching and special needs teaching); (ii) individual teacher data from matched employer-employee data, which enable us to study teacher characteristics; (iii) student-level outcome data from administrative registers, including standardized tests, end-of-compulsory-school exam scores and teacher grades, and an early measure of progression in upper secondary school; and finally, (iv) school-level responses from an annual national student survey , allowing us to study school environment attributes.

Table 1 shows key descriptive statistics for our main student-level estimation sample, separately for the pre-funding (2009-2012) and the funding (2013-2016) periods. Here year refers to the first semester of the school year, and students are indexed by their first year in lower secondary (e.g., 2009 represents the school year 2009/2010 and the students in grade 8 this school year). We include the 2009-2019 cohorts in the analysis, in total about 618,000 students (about 56,000 students per cohort). 169,000 students, or 27%, attended treatment schools that received extra funding during 2013-2016. The treated schools have about one standard deviation lower GPA than other schools

[4] We define treatment from the forcing variables in our data, which may not correspond exactly to the data used by the ministry when they allocated extra funds. For all but one school in our sample, the expected and recorded numbers of teachers match. This school, with 257 students, did not receive extra teachers, despite being just below the GPA cutoff (and well above the student-teacher ratio cutoff) in our data.

Figure 1. Pre-reform GPAs and student-teacher ratios of treated and control schools



Note: The figure shows the schools by student-teacher ratio and GPA in the pre-reform year 2011/2012. These forcing variables are constructed from student and school registers. The red dotted lines mark the two cut-offs. The markers indicate whether the schools received extra teachers (from a separate data source).

and about one standard deviation more students per teacher in the pre-treatment year of 2011 (panel A in Table 1). This corresponds to a student-teacher ratio in regular teaching of 22 (19) in the treated (non-treated) schools. Furthermore, the treated schools are on average significantly larger, with about 341 students compared to 279 students for the non-treated schools. GPA levels, student-teacher ratios, and school size all vary more among the control schools than among the treatment schools.

In panel B of Table 1, we report student characteristics. While the sex composition is balanced, the treated schools have fewer students with at least one college-educated parent, more fathers with below-median earnings, and a larger fraction of students with two immigrant parents. These differences are expected since the funding targeted schools that have below-average GPAs. However, the differences are mostly modest, with 6 percentage points more students having at least one college-educated parent and the same difference for above-median earnings in the untreated schools. Similarly, the average entry test score differential in the 8th grade between treated and untreated schools changes from -0.13 SD (pre) to -0.1 SD (post).

In Panel C we present mean student-level outcomes. The 9th grade test is similar to the 8th grade test and taken early after just over 12 months in lower secondary school. For the analysis we normalize the 9th grade test scores, exam scores, and teacher grades within each cohort. At the end of lower secondary school, grade 10 after three years, the students sit one written anonymously graded exam in either Norwegian, English or mathematics. The student is also assigned grades in about 13 different subjects by their classroom teachers. Our final student-level outcome is on-time completion of year two of upper secondary. Unlike lower secondary, while almost all student enroll in upper secondary directly after completing lower secondary, upper secondary is not compulsory and progression not automatic, the students need to pass all subjects. We observe whether the student has completed year two of upper secondary within two years of completing lower secondary. For the untreated schools, there is a mix of positive and negative changes in student outcomes from pre- to post-treatment cohorts. All four outcomes improve for the treatment schools from pre- to post-treatment cohorts, potentially indicating positive intervention effects. However, these improvements might also simply reflect mean reversion, as the schools were initially selected due to their poor performance in 2011 (Chay et al., 2005). In the following, we investigate whether these improvements can be interpreted as effects of the intervention.

Table 1. Descriptive statistics

	Untreated schools		Treated schools	
	Pre-years	Post-year	Pre-years	Post-years
A. School characteristics	2011		2011	
GPA (de-meaned and standardized)	0.202		-0.743	
	(1.013)		(0.544)	
Student-teacher ratio (de-meaned and standardized)	-0.245		0.734	
	(0.999)		(0.570)	
Student-teacher ratio	19.0		22.4	
	(3.56)		(2.21)	
School size (# of students)	279		341	
	(139)		(102)	
B. Student characteristics	2009-2012	2013-2016	2009-2012	2013-2016
Female students (share)	0.488	0.491	0.487	0.487
	(0.500)	(0.500)	(0.500)	(0.500)
No parent with higher education (share)	0.482	0.429	0.543	0.488
	(0.500)	(0.495)	(0.498)	(0.500)
Father has below median income (share)	0.488	0.485	0.533	0.540
	(0.500)	(0.500)	(0.499)	(0.498)
Two foreign-born parents (share)	0.084	0.109	0.147	0.177
	(0.278)	(0.312)	(0.354)	(0.382)
Entry test score 8th grade	0.022	0.016	-0.114	-0.082
	(0.920)	(0.924)	(0.920)	(0.919)
C. Student outcomes				
9th grade test score	0.024	0.021	-0.100	-0.074
	(0.917)	(0.922)	(0.928)	(0.928)
Written exam score 10th grade	0.053	0.071	-0.050	-0.009
	(0.989)	(0.973)	(0.985)	(0.972)
Teacher assessment grade 10th grade	0.044	0.031	-0.100	-0.067
	(0.985)	(0.984)	(1.012)	(1.011)
Completed 2.year of upper secondary school on time	0.781	0.816	0.759	0.797
	(0.413)	(0.388)	(0.428)	(0.402)
Number of students	168 305	160 132	63 135	60 712

Note: Years refer to year of school register data and 8th grade test score, and the first semester of the school year. Standard deviation in parentheses. School-level observations are weighted by number of students. The pre-treatment school characteristics are those used to assign schools to treatment, and are based solely on 2011 data. Entry test score is the average of reading, numeracy, and English, which we normalize within each cohort to have zero mean and a standard deviation of one. On-time completion of second year in upper secondary is observed two years after completion of lower secondary, and is only observed for 2015 and earlier cohorts, other outcomes are observed for 93-99% of students (most missing for written exam score and teacher grades). Test scores, exam scores and teacher grades are standardized within year in the total student population.

3 Empirical strategy

We use two distinct approaches to identify the effects of the policy: a local RD model and what we call a global difference-in-RD model. Both identification strategies exploit the strict funding assignment to estimate the impact of hiring more teachers. However, the local RD specification focuses on schools near the specific cutoffs, while the global difference-in-RD specification includes all schools, irrespective of their distances from the cutoffs. As a result, the two approaches differ concerning identifying assumptions and treatment effects estimated. We begin by describing the local RD specification before turning to our preferred difference-in-RD specification below.

3.1 Local regression discontinuity design

In the local RD specification, we exploit the discontinuous change in the school´s probability of receiving extra funding for teachers near the student-teacher ratio and GPA margins to estimate local average treatment effects (LATE). Schools received additional funding if they had a student-teacher ratio (STR) above 20 and a grade point average (GPA) below the national mean in 2011 (i.e., the school year 2011/2012, following the notation from the previous section). These two necessary conditions (cutoffs) place each school in one of the four quadrants of Figure 1 and we label them treatment schools (southeast), GPA placebo schools (southwest), STR placebo schools (northeast), and control schools (northwest). The strict funding assignment offers two distinct margins for evaluating treatment effects and two placebo margins that can be used to validate the design.

In local RD estimation, there is a trade-off between unbiasedness (improved by a small bandwidth) and precision (improved by a wider bandwidth and more data). With two forcing variables, there are several choices to be made regarding how to summarize the effects and analyze the data that add to the questions of optimal bandwidths and the functional form of the running variable(s). For instance, separate analyses can be run for the two different margins, and this can be done either parametrically or non-parametrically. Alternatively, both cutoffs can be used simultaneously (Cattaneo et al., 2020).

We estimate effects separately for the two different cutoffs: the STR and the GPA margin. Along these dimensions, the sample is restricted to +/- 0.5 standard deviation in the separate RD analyses. We use the rdrobust and rdplot packages (Calonico et al., 2014) to estimate local linear regressions and bias-corrected confidence intervals and to plot the results.[5] For the STR margin, we limit the local RD analyses to schools with a below-average GPA (i.e., treatment and GPA placebo schools) and compare schools near the STR threshold cutoff. Similarly, for the GPA margin, the sample is restricted to schools with an above-average student-teacher ratio (i.e., treatment and STR placebo schools), and we compare the schools near the GPA threshold cutoff.

The local RD rests on the assumption that there is no strategic sorting near the cutoffs. Importantly, since the extra funding was decided in Parliament without a long public debate in advance, the historical data used in the allocation formula can be assumed exogenous. We find

[5] Rather than using the RMSE optimal bandwidth selection of the rdrobust and rdplot packages, we will use a fixed bandwidth of 0.5 SD for the RD estimation. We found that the RMSE optimal bandwidths (typically 0.2-0.3 SD) produced implausibly strong gradients around the cut-offs and correspondingly implausibly large and imprecise estimates, cf. Figures 2 and A.2.

it highly unlikely that schools were able to manipulate their GPA or the student-teacher ratio that determined eligibility. Furthermore, in Figure A.1 in the Appendix, we also demonstrate that there is no indication of bunching around any of the cutoffs, supporting the idea that schools could not manipulate their treatment status. Additionally, because the funding assignment was defined based on two dimensions, we also have two placebo margins that can be used to validate the design. We can compare schools near the GPA threshold with below-average student-teacher ratios (i.e., GPA placebo schools and control schools) and schools near the student-teacher ratio with above-average GPA (STR placebo schools and control schools). As shown in the result section, these placebo tests provide evidence that our local RD is unbiased.

While providing a credible approach to identifying treatment effects, the statistical power to detect small effect sizes with the local RD specification is a concern, and treatment effects for schools further from the cutoffs may differ from the local effects identified by outcomes near the cutoffs. Motivated by these concern, we present an alternative parametric "global" specification in the next section that provides greater precision than the local RD.

3.2 Global difference-in-RD design

Our preferred model is what we term a global (parametric) difference-in-RD specification in which both cutoffs and all schools, irrespective of their distances to the cutoffs, are used to identify the effects of the intervention. This design allows us to use both cutoffs simultaneously and provide a more precise estimate than the local RD while at the same time having several placebo dimensions that can be used to evaluate the design.

To simplify the presentation of the model, let z_S and z_G be the forcing variables on the student-teacher ratio and GPA margins, measuring the distance to cutoffs. We estimate

$$y_t = \delta_t(Z_G Z_S) + \gamma_{1t} Z_G + \gamma_{2t} Z_S + \eta_{1t} z_G + \eta_{2t} z_S + \eta_{11t} z_G^2 + \eta_{22t} z_S^2 + \eta_{12t} z_G z_S + x_t \beta + \epsilon_t, \quad (1)$$

where y_t denotes the outcome of interest for cohort t (defined by entry year in grade 8), and Z_G and Z_S are dummies for $Z_G := z_G > 0$ and $Z_S := z_S > 0$ (i.e., crossing the separate cutoff thresholds). Treated schools have $z_S > 0$ and $z_G > 0$ (both measured in the pre-treatment year 2011), denoted by $Z_G Z_S$ in Equation (1). Finally, x_t is a vector of control variables, while ϵ_t is the residual.

The main coefficient of interest is δ_t, which shows the effects of being a student in cohort t in a school above both cutoffs. This difference-in-RD design is similar to a two-by-two difference-in-difference setup, where the Z_G and Z_S variables control for potential (cohort-specific) differences in outcomes around the two cutoffs in untreated schools. However, unlike in a simple two-by-two setup, we also observe distances to the two cutoffs (i.e., z_S and z_G), and include quadratic distance controls, as shown in Equation (1). Under standard RD regularity assumptions and absent treatment, sufficiently flexible distance controls will control for differences in potential outcomes around the cutoffs, making the controls for above/below the two cutoffs superfluous. In this case, $\gamma_1 = \gamma_2 = 0$ and the inclusion of Z_G and Z_S variables is not necessary for consistent estimation of δ_t. By way of contrast, suppose there are discontinuities around the cutoffs in the absence of treatment or that the distance controls are not sufficiently flexible to capture differences around the cutoffs. In that case, the inclusion of Z_G and Z_S accounts for such differences and is necessary to identify δ_t.

Additionally, the Z_G and Z_S variables serve as a placebo test. Evidence that $\gamma_1 = \gamma_2 = 0$ indicates that the controls for the distance to the cutoffs are sufficient to control for differences between treated and untreated schools and further strengthens the credibility of the effect estimate δ_t. This placebo test builds upon the reasoning that, in the absence of the treatment, crossing both cutoffs should not affect the outcome if crossing either of the two cutoffs has no effect.[6]

As indicated by the cohort subscripts on the δ_t, γ_t, and η_t coefficients in Equation (1), all inference is based on variation within the cohort, comparing outcomes of treated and untreated schools, conditional on the forcing variables. For treated cohorts, δ_t provides our main policy effect estimates. For later cohorts who attended grades 8-10 after the extra funding was terminated, δ_t provides estimates of lasting effects on the schools, potentially including any local continuation of the extra teachers using other funding. For cohorts in grades 8 to 10 before the reform, δ_t shows whether there were pre-existing differences between treatment and control schools, providing us with another placebo check. In addition, we run placebo regressions where we test for discontinuities in background characteristics such as entry test scores and parental education.

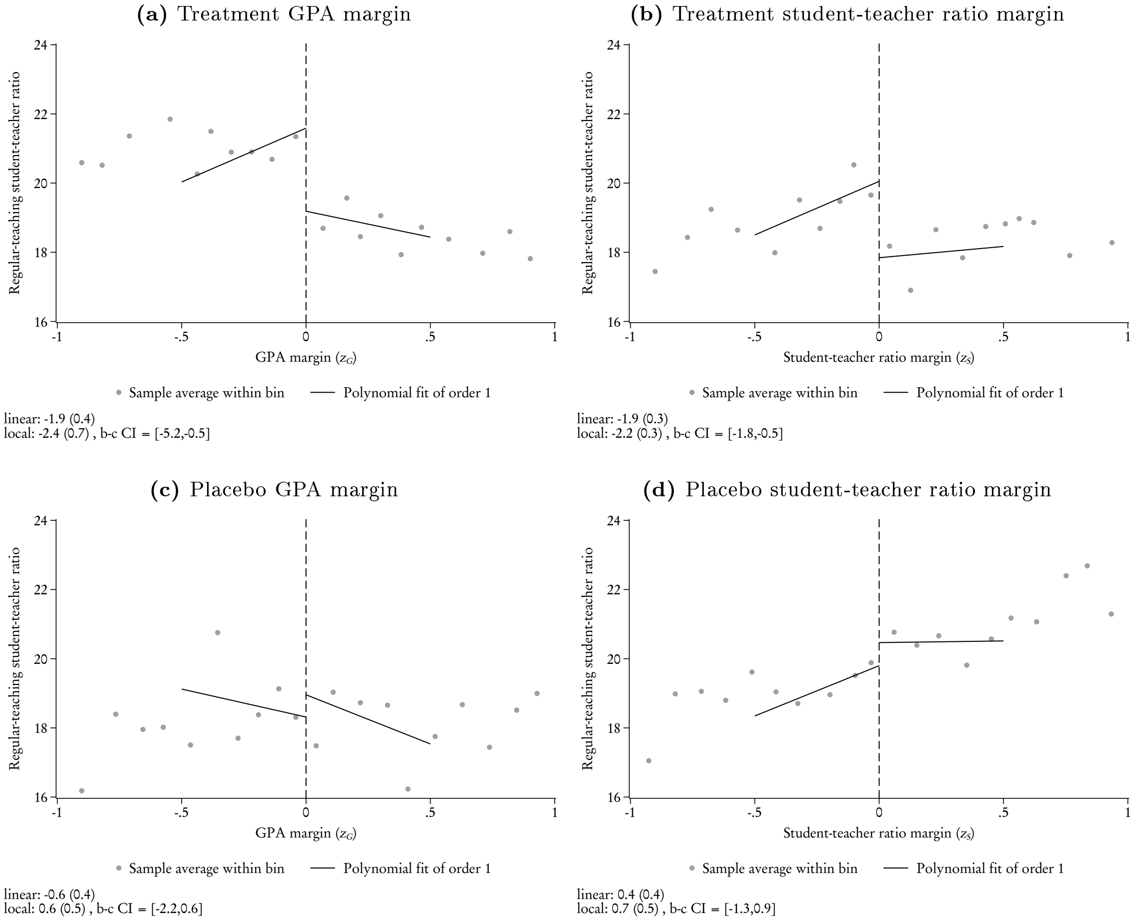
The control variables in x_t include gender, year fixed effects, a cubic in the 8th-grade test score, and parental education (dummies for seven levels of education for each parent). Control variables are included to increase precision, and results without controls are very similar (as shown in the Appendix). Because treatment is at the cohort-school level, standard errors will be clustered at the school level when we study individual-level outcomes. The results will be summarized in coefficient plots with confidence intervals.

The local and global RD estimates differ in terms of precision, identification, and, if there is effect heterogeneity, estimated treatment effects. Starting with precision, the global difference-in-RD estimates are substantially more precise than the local RD estimates because the former uses data from all schools and includes controls that reduce the residual variation. However, this gain in precision comes at the cost of moving away from the well-identified LATEs around the cutoffs. Despite this, as discussed above, the difference-in-RD specification provides placebo tests both from the pre-treatment years and from the placebo margins during the treatment years. These placebo tests allow us to convincingly evaluate the credibility of the global RD design. Finally, unlike the local RD estimates, the difference-in-RD estimate δ_t does not correspond to a LATE around the cutoff – rather, it represents an average effect across all treated schools. Consequently, differences between the local and global RD estimates may reflect treatment effect heterogeneity among schools at various distances to the cutoffs.[7]

[6] The placebo tests also serve as test of mean reversion. Mean reversion will, on average, contribute to improved outcomes in schools assigned to treatment because of low GPA pre-assignment (Chay et al., 2005), but will typically not bias RD estimates since the degree of mean reversion is similar on both sides of the funding threshold. Because we have both treated and untreated schools with low GPA pre-assignment, any mean reversion not handled by the parametric RD will appear in the placebo tests.

[7] If there is treatment effect heterogeneity that correlates with the forcing variables z_G and z_S, this heterogeneity will impact the estimates of η_{1t}, η_{2t}, and δ_t, such that δ_t may not reflect an average treatment effect. As a robustness check, we have first residualized y_t with respect to the forcing variables, using only the non-treated schools, and then estimated δ_t, γ_{1t}, and γ_{2t} from the residualized data (not controlling for the forcing variables in the second regression). The results are very similar to our main results.

Figure 2. Local RD estimates of the effect on student-teacher ratio



(a) Treatment GPA margin

(b) Treatment student-teacher ratio margin

(c) Placebo GPA margin

(d) Placebo student-teacher ratio margin

(a) linear: -1.9 (0.4)
local: -2.4 (0.7) , b-c CI = [-5.2,-0.5]

(b) linear: -1.9 (0.3)
local: -2.2 (0.3) , b-c CI = [-1.8,-0.5]

(c) linear: -0.6 (0.4)
local: 0.6 (0.5) , b-c CI = [-2.2,0.6]

(d) linear: 0.4 (0.4)
local: 0.7 (0.5) , b-c CI = [-1.3,0.9]

Note: The graphs show RD estimates. Data are from the years 2013-2016 and the outcome is the student-teacher ratio for regular teaching. Figure notes show coefficients and standard errors from linear and local linear regressions. All analyses use school-level data and student weights. The lines show the local linear regressions and are estimated using Calonico et al. (2014), with triangular weights and a fixed bandwidth of 0.5. Bias-corrected confidence intervals (estimated using higher-order polynomial) in brackets. Bins are quantile-based.

4 School resources

The policy funded new teaching positions in targeted schools over a period of four years (2013-2016). We start by examining how it affected the student-teacher ratio. We first present results for the two treatment and placebo margins in Figure 2 using the RD design. For the treatment GPA margin (with $z_S > 0$), there is a clear drop in the student-teacher ratio around the cutoff from about 21.5 to 19 (Panel (a) of Figure 2). For the treatment student-teacher ratio margin (with $z_G > 0$), the linear, local linear, and bias-corrected local linear estimates are all similar and statistically significant with a reduction of about two students per teacher (Panel (b)). As in the case of the GPA margin, there is a drop of about two students per teacher, from 20 to 18, and all estimates are of the same magnitude and statistically significant.

In sub-figures (c) and (d) of Figure 2, we show the corresponding results for the two placebo margins (i.e., with $z_S < 0$ and $z_G < 0$, respectively). Consistent with our expectations, we find zero effects on the student-teacher ratio from crossing the two separate cutoffs. All placebo estimates are small and statistically insignificant.

The estimates of our preferred difference-in-RD specification in Equation (1) are presented in Figure 3. The figure displays the point estimates and 95% confidence intervals for separate school cohorts in treatment schools ($Z_G Z_S$,δ; blue circle symbol), GPA margin placebo schools (i.e., Z_G, γ_1; green diamond symbol), and placebo schools at the student-teacher ratio margin (i.e., Z_S, γ_2; red triangle symbol). All estimates are within-year, relative to the group of schools with high GPA and a low student-teacher ratio, and conditional on the parametric specification in Equation (1). Note that this parametric specification controls for the forcing variables as measured in 2011, which means that the 2011 estimates will be zero as the outcome is the same as one of the forcing variables in this year. The treatment coefficients during the shaded period in the graph (2013-2016) can be interpreted as effects of the funding.

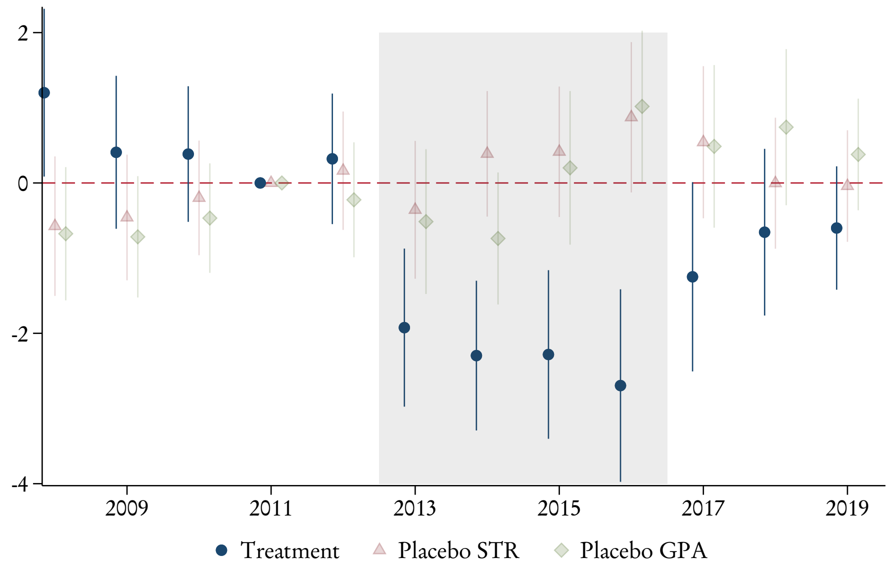
Figure 3 shows that the policy reduced the class size during the treatment years (blue circles in the shaded area of the figure). The average effect over the four treatment years is a reduction of 2.3 in the student-teacher ratio, which is substantial compared to the treated school average of 22.4 before the reform. In Figure A.3 in the Appendix, we find the same result using logs (11%). Moreover, in Appendix Figure A.4, we show that the policy increased the number of teachers by about 3.9 teachers (17%) in treated schools, closely corresponding to the number of teachers funded. Thus, there is no evidence of municipalities' funding being crowded out or substitution to other schools (Section 7 discusses substitution effects in more detail). The evidence clearly shows that the funding policy had the intended effect on teacher input.[8] While there are no significant effects on student-teacher ratios after the termination of the reform, there are indications of a gradual reversion to control school levels. This is consistent with reports that some municipalities continued with a reduced student-teacher ratio funded by other sources.

We check the validity of our design using three distinct placebo dimensions as well as the pre-treatment balancing tests discussed in Section 5, all of which support our identification strategy. The first placebo dimension is treatment school coefficients before the funding period (2009-2012). Changes in the student-teacher ratio (or other outcomes) before the funding period would suggest trends within treatment schools that may bias our effect estimates. We see no indications of significant differences in schools that later become treatment schools.

The within cohort placebo dimensions are the non-funded schools with low GPA and high student-teacher ratio, respectively. Since these placebo schools did not receive funding during the period, they can be used to rule out other changes concurrent with the funding policy that presumably would have affected all schools with either low GPAs or high student-teacher ratios. These coefficients seem to trend over the reform period and are close to significant in 2016. However, as discussed in Section 3, these coefficients do not have to be zero for effect estimates to be valid. Over time, the forcing variables may become less predictive of later outcomes, and controls for low initial GPA and high student-teacher ratios may become more important for correct inference. Nevertheless, there are no significant differences on either placebo margin in

[8] This was expected, but in light of another Norwegian education policy of resources to primary schools not leading to lower student-teacher ratios (Reiling et al., 2021), not obvious.

Figure 3. Parametric difference-in-RD estimates of effects on the student-teacher ratio



sample mean (SD) = 18.0 (3.8)
pooled 2013-2016 estimate = -2.3** (0.3)
p-value joint test of 2013-2016 placebos = 0.1350

Note: The graph shows estimates and confidence intervals from estimating eq. (1). The outcome is the student-teacher ratio. The different series correspond to treatment effects, δ in eq. (1), and placebo effects from the non-treatment margins, γ in eq. (1). The treatment period is shaded. The text below the figure reports the sample mean and standard deviation of the outcome, estimated effect and standard errors for a pooled analysis of the treatment years, and the p-value of a joint test of all placebo effects (γ) for all treatment years. The regression uses school-level data with student weights and robust standard errors.

any year or in total over both margins and all reform years (p-value of 0.135), and the placebo estimates are small compared to the effect estimate. These results indicate that the parametric specification of the forcing variables is sufficient to control for between-school differences around the cutoffs and strengthens the credibility of the estimated treatment effects.

In Appendix Table A.3, we investigate the robustness and compare estimates from alternative specifications, including Equation (1) with global linear controls, as well as local specifications with either linear or quadratic controls. The local specifications are similar to those in Figure 2, based on observations within 0.5 SD of either cutoff using triangular weights based on distance to the closest cutoff. Although the estimated effect on student-teacher ratio is slightly larger with the global linear specification, the effect is not significantly different across specifications. In the global linear specification, the coefficient for the STR placebo margin is also significant, suggesting that the quadratic terms are needed to fully capture school heterogeneity. The estimated coefficients from the two local specifications are very similar to those of our preferred specification, but the standard errors are larger.

In sum, we conclude that there is strong and robust evidence of a substantial reduction in the student-teacher ratio, from about 22 to 20 students per teacher in regular instruction.

5 Academic outcomes

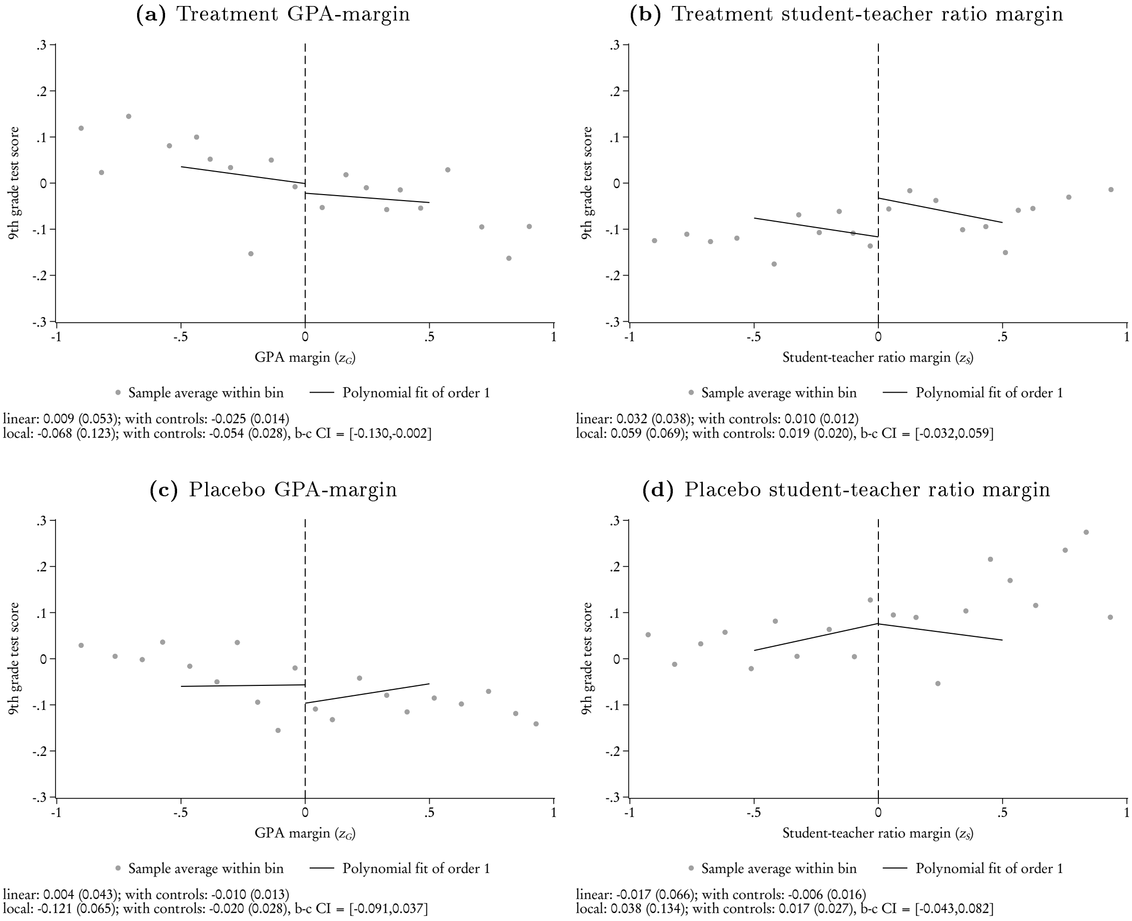
The main objective of the intervention was to improve student outcomes. Since the additional funding lasted for four yours, exposure years differs across cohorts and outcomes. We start by studying the effects on the standardized test scores in grade 9, when students had been exposed to the treatment for about one school year. Figure 4 shows results from school-level RD analyses of the two margins, similar to the school resources analysis in Section 4. If the extra teachers improved student outcomes, we should see a jump in test scores at the two margins. There is no indication of any test score discontinuity in neither treatment nor placebo margins. The estimated effects are small in all sub-figures and never statistically significant. However, the RD estimates are not very precise, and in several cases, we cannot reject intervention effects of 5% of a standard deviation.

Figure 5(a) shows the effects on students' 9th-grade standardized test scores from our main specification in Equation (1). One year with a lower student-teacher ratio did not improve the treated students´ scores on the national standardized test. The average estimated effect across the four treated cohorts is 0.6% of a standard deviation, with a confidence interval from -1.6% to 2.8%. The results are similar without control variables (see Appendix Figure A.8), but the precision is much lower. All placebo test estimates are small and insignificant. The pre-reform placebo effects are similar in size to the reform effects. Moreover, the within-year placebo estimates are close to zero, thus providing additional evidence of the validity of our design. All in all, the effect of more teachers on the 9th-grade test scores is minor at best.

Compositional changes within treatment schools represent a potential concern. We examine this by studying placebo effects on students' entry test scores in lower secondary schools as well as on parental characteristics. Since funding in lower secondary schools cannot influence test scores at entry, any significant placebo effects would indicate compositional changes correlated with the treatment. Although we control for students' entry test scores when estimating the effect on 9th grade test scores, which accounts for this exact source of bias, any within-school changes in students' academic abilities correlated with treatment would be worrying since it would suggest that other unobserved factors were co-occurring with our treatment. Reassuringly, there are no discontinuities in entry test scores nor in parental earnings, parental education, or immigrant background in treatment schools before, during, or after the funding period, as seen in Appendix Figure A.9. Nor are there any systematic differences in any of the placebo schools. All in all, the placebo and balance tests make us confident that composition bias is unlikely.

The precise zero effect on 9th grade test scores strongly restricts potential significant effects for subgroups. Any effect of the extra funding is either small, limited to a small group of students, or counteracted by a negative effect for other students. Previous research has found that it is primarily disadvantaged students who benefit from extra funding (Jackson, 2020). In Appendix Figure A.10, we investigate heterogeneous effects by individual (gender, earlier test scores) and parental characteristics (immigrant background, income, and education). We do not find significant effects for any group. Indeed, we can reject effects larger than 4% of a standard deviation for most groups. The single exception is children of immigrants, for whom we can only reject effects larger than 6%. This mostly reflects lower precision for this group (which is smaller than the others studied), since the point estimate of less than 2% hardly indicates any large effect. In Appendix Figure

Figure 4. Local RD estimates of effects on 9th grade test scores



(a) Treatment GPA-margin

linear: 0.009 (0.053); with controls: -0.025 (0.014)
local: -0.068 (0.123); with controls: -0.054 (0.028), b-c CI = [-0.130,-0.002]

(b) Treatment student-teacher ratio margin

linear: 0.032 (0.038); with controls: 0.010 (0.012)
local: 0.059 (0.069); with controls: 0.019 (0.020), b-c CI = [-0.032,0.059]

(c) Placebo GPA-margin

linear: 0.004 (0.043); with controls: -0.010 (0.013)
local: -0.121 (0.065); with controls: -0.020 (0.028), b-c CI = [-0.091,0.037]

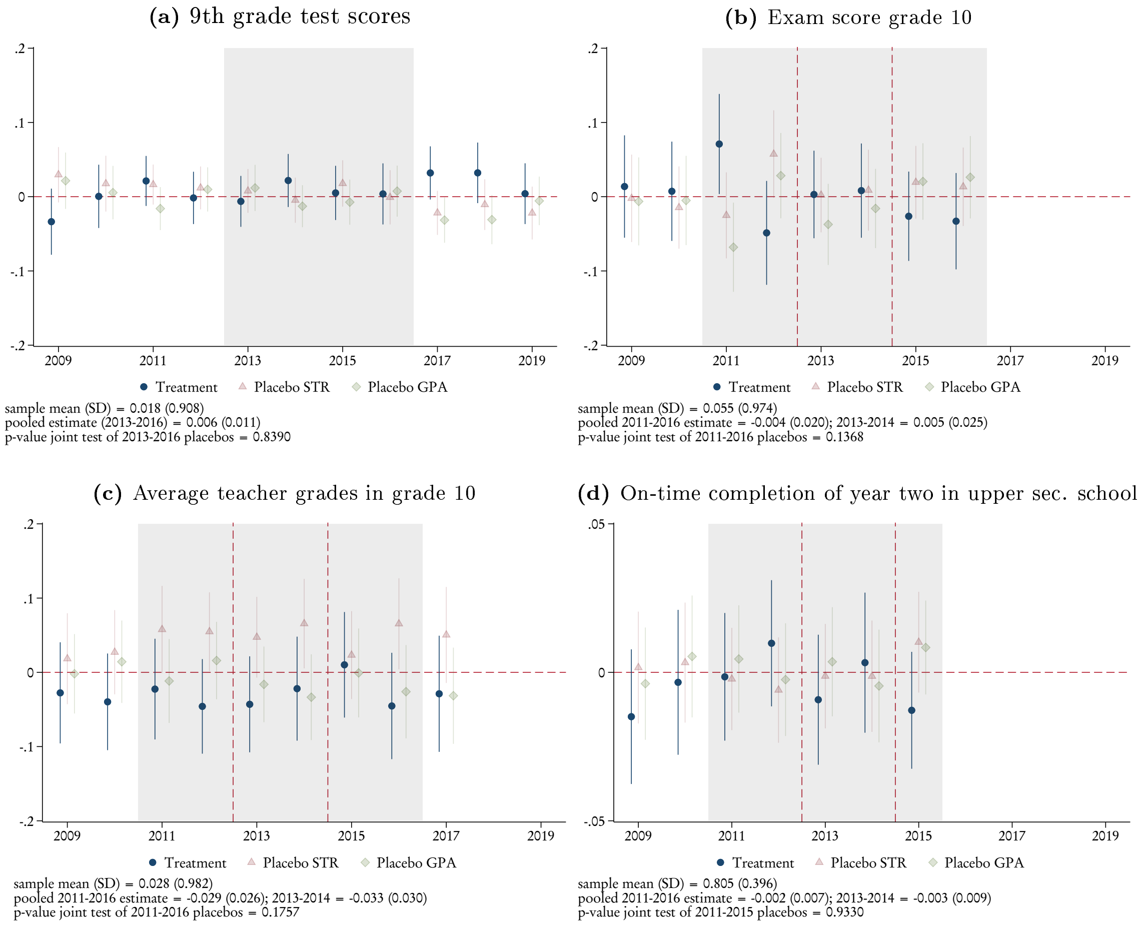
(d) Placebo student-teacher ratio margin

9th grade test score
.3
.2
.1
0
-.1
-.2
-.3
-1
-.5
0
.5
1
Student-teacher ratio margin (z_S)
● Sample average within bin
—— Polynomial fit of order 1

linear: -0.017 (0.066); with controls: -0.006 (0.016)
local: 0.038 (0.134); with controls: 0.017 (0.027), b-c CI = [-0.043,0.082]

Note: The graphs show RD estimates for the average score from the 9th grade test. The data are provided by students sitting the grade 9 test in 2013-2016. The graph is based on school-level data and student weights. The lines show the local linear regressions and are estimated using Calonico et al. (2014), with no additional control variables, triangular weights and a fixed bandwidth of 0.5. Bins are quantile-based. Figure notes show coefficients and standard errors from student-level linear and local regressions. Student controls in the linear and local regressions include gender, year dummies, a cubic in the 8th grade test score, and parental education. Bias-corrected confidence interval (estimated using higher-order polynomial) in brackets.

Figure 5. Parametric difference-in-RD estimates of effects on academic outcomes



(a) 9th grade test scores

(b) Exam score grade 10

(c) Average teacher grades in grade 10

(d) On-time completion of year two in upper sec. school

sample mean (SD) = 0.018 (0.908)
pooled estimate (2013-2016) = 0.006 (0.011)
p-value joint test of 2013-2016 placebos = 0.8390

sample mean (SD) = 0.055 (0.974)
pooled 2011-2016 estimate = -0.004 (0.020); 2013-2014 = 0.005 (0.025)
p-value joint test of 2011-2016 placebos = 0.1368

sample mean (SD) = 0.028 (0.982)
pooled 2011-2016 estimate = -0.029 (0.026); 2013-2014 = -0.033 (0.030)
p-value joint test of 2011-2016 placebos = 0.1757

sample mean (SD) = 0.805 (0.396)
pooled 2011-2016 estimate = -0.002 (0.007); 2013-2014 = -0.003 (0.009)
p-value joint test of 2011-2015 placebos = 0.9330

Note: The graphs show estimates and confidence intervals from estimating eq. (1). Outcomes are a) (standardized) 9th grade test scores, b) (standardized) exam score, c) (standardized) teacher grades, and d) completion of year two of high school. Control variables are gender, age, year fixed effects, a cubic in the 8th grade test score, and parental education. The different series correspond to treatment effects, δ in eq. (1), and placebo effects from the non-treatment margins, γ in eq. (1). The X-axis is the year of the 8th grade test, treated cohorts are shaded. In sub-figures b-d the dashed vertical lines indicate cohorts treated for three years. The figure notes show the sample mean and standard deviation of the outcome, estimated effect and standard errors for a pooled analysis of the treatment years and the p-value of a joint test of all placebo effects (γ) for all treatment years. The regression uses student-level data and clusters standard errors at school level.

A.11 we investigate heterogeneity by school characteristics and find no significant effect for most types of schools, with the exception of marginally significant effects in schools with lower average 8th grade test scores. In all, we findno evidence for heterogeneous effects.

Treatment school students had been exposed to a lower student-teacher ratio for just over one school year when they took the 9th grade test. Since a longer exposure may have a more substantial impact, we also estimate the effects at grade 10. By the end of compulsory schooling (grade 10), the 2011 and 2016 cohorts had been exposed to one year of extra funding, the 2012 and 2015 cohorts to two years, and the 2013 and 2014 cohorts to three years. Panel (b) of Figure 5 shows the effects on externally graded exam scores at the end of compulsory schooling.[9] The average effect across all treated cohorts is -0.4% of a standard deviation. According to a dose-response rationale, being exposed to more resources over a longer period should have a larger impact (see Jackson and Mackevicius, 2021). However, the average impact for school cohorts exposed for three years to extra funding, delimited with dashed vertical lines, is only 0.5% of a standard deviation and insignificant.[10] While Jackson and Mackevicius (2021) find a linear dose-respone, the precision of effect estimate for the fully treated 2013 and 2014 cohorts is slightly lower than the precision of the estimate using all cohorts. Thus, the fully treated cohorts provide the best opportunity to find any treatment effect if it exists and will therefore be our main focus.

We also investigate the effects on a broader set of academic outcomes. To the extent that non-cognitive skills impact longer-term outcomes, we would expect that such effects show up in teacher-graded tests and school dropout. Specifically, since the funding did not improve the 9th grade standardized test scores nor 10th grade exam scores, any effects on 10th grade teacher-assessed grades and upper secondary school completion would likely have been caused by non-cognitive skills (or effects on teachers' grading practices). We find no significant effect on teacher-assessed grades. The estimated average effect across all treated cohorts is -2.8% of a standard deviation for teacher-assessed grades, and the impact for those exposed for three years is -3.3% of a standard deviation. Finally, there is no significant effect for on-time completion of the second year of upper secondary school, with the 95% confidence interval ranging, in terms of percentage points, from a reduction of 1.6 to an increase of 1.2.[11]

Appendix Figure A.8 shows that the results are similar without controls. As for the effect on the student-teacher ratio, the estimated effects on academic outcomes are robust with respect to

[9] RD graphs similar to Figure 4 for the outcomes in panels (b)-(d) of Figure 5 are included in the Appendix, as Appendix Figures A.5-A.7.

[10] While the point estimate for 2011 is positive and significant, there is no clear pattern for the whole set of estimates, and a joint test of all the effects for the period 2011-2016 does not reject the null of zero effects (p-value = 0.136). Furthermore, the positive effect estimate in 2011 is also accompanied by negative estimates at both placebo margins, reducing the total difference between treated schools and schools with high GPAs and low STRs. Finally, we did not find any effect on this cohort's 9th grade test scores (discussed above), concurrent teacher grades, or on later school dropout (both discussed below). Thus, we interpret the 2011 effect as spuriously significant and not indicating any real effect on this cohort. Given the number of hypotheses tested, one or more spuriously significant result is not unexpected, and the 2011 estimate is only just significant (p-value = 0.039).

[11] We are not able to observe the treated cohorts complete upper secondary school, which nominally takes from 3-4.5 years (depending on the track) and which is customarily measured five years after completion of lower secondary school (or enrolling in upper secondary school). However, based on earlier cohorts, completion of the second year strongly predicts eventual completion of upper secondary school. At about 80%, the share of students completing the second year on time is similar to the share completing upper secondary school within five years. Students completing the second year on time have a 50 percentage points higher probability of completing upper secondary school in Norway. Conditioning on results from lower secondary school, gender, and parental education reduces this difference to 34 percentage points. In Figure A.12, we investigate on-time enrollment in the third year, and find insignificant negative effects.

the choice of specification, see Appendix Tables A.5-A.7 where we investigate the robustness of the effect estimate for the fully treated cohorts.

6 School environment

This section examines whether the funding affected students in ways not captured by academic outcomes. In a nationwide survey of Norwegian 10th graders, students respond anonymously on subjects such as well-being in school, teacher support, and bullying. We construct an index based on aggregate student responses from 11 sub-indices to test for any effect on the school environment. The students score each item on a scale from one to five, and the sub-indices represent school-level averages of students' average scores across a small number of related questions. Except for bullying, higher values imply a better school environment.[12] To construct the index, we scale each sub-index by dividing by the student-level standard deviation and average the scaled sub-indices (bullying is rescaled as 1 - sub-index before averaging). A higher value implies a better outcome in all sub-indices and in the overall index).[13]

Unlike in analyses of academic outcomes, we find evidence suggesting that the extra teachers improved the school environment index by about 5% of a student-level standard deviation, as shown by Figure 6. While all single-year effect estimates are insignificant, they are all positive, and the average effects over the 2011-2016 cohorts (exposed for one, two, or three years to extra teachers) and 2013-2014 cohorts (exposed for three years to extra teachers) are significant at the 5% level.[14]

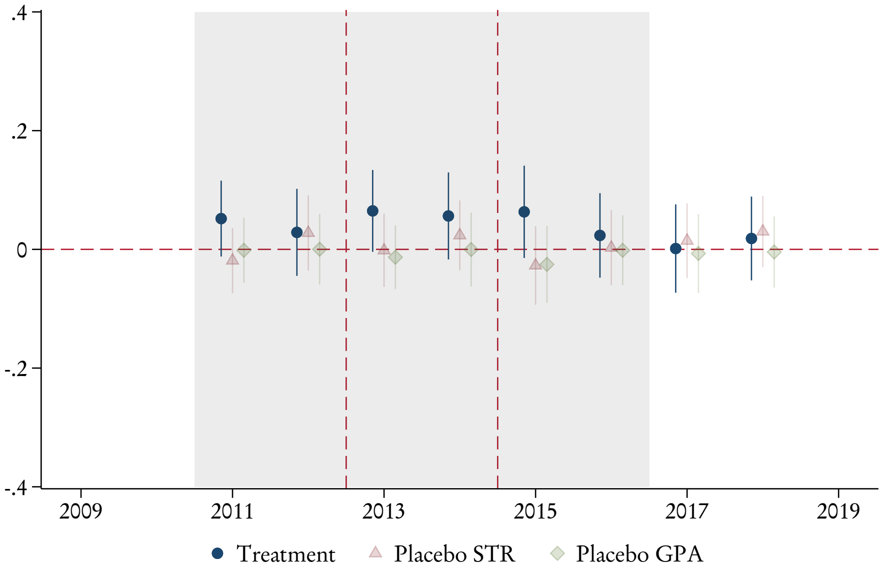
The validity of the design and the credibility of the positive treatment effects are supported by the small and insignificant placebo effects in Figure 6. Since the student survey was redesigned in 2012, we do not have data for all sub-indices in all pre-treatment years. However, there are no indications of pre-treatment differences for the sub-indices that are covered throughout (Figures A.14 and A.15). Moreover, the school environment improvements among treated schools disappear once the funding ends, as shown by the precisely estimated zero treatment coefficients for 2017 and 2018. Notably, the effects are already reduced for the 2016 cohort. When the 2016 cohort answered the survey in the fall of 2018, the funding had terminated more than a year before, and there was little difference remaining in student-teacher ratios (see Figure 3). In contrast, the 2015 cohort answered the survey a few months into the first school year after the end of the reform, when differences in student-teacher ratios were still large, and when some schools may still have had extra teachers (see Figure 3). While the estimates are not sufficiently precise to reject equality of the yearly effects, this cross-cohort pattern suggests that the school environment effects are relatively short-lived.

[12] For example, the students are asked “Do you enjoy school?” with possible answers “Not at all”, “Not very much”, “Somewhat”, “Quite well”, and “Very well”, and asked “Do you feel that your teachers care about you?” with possible answers “None”, “Only one”, “A few”, “Most of them”, and “All”.

[13] The student-level standard deviation ranges from 0.7-1.0 in the original one to five units, meaning effects in original units are slightly larger than the standardized effects we report.

[14] Corresponding estimates for each of the 11 different sub-indices are presented in Figures A.14 and A.15. These constitute all published sub-indices with two exceptions: share bullied, which is a transformation of the bullying sub-index, and which is measured as a share of the students, and support from parents, which does not reflect the classroom environment, and which we will return to in Section 7. RD graphs investigating effects on the overall index on all margins are presented in Appendix Figure A.13.

Figure 6. Effects on the school environment index



sample mean (SD) = 4.572 (0.215)
pooled 2011-2016 estimate = 0.048** (0.015); 2013-2014 = 0.061** (0.026)
p-value joint test of 2011-2016 placebos = 0.9948

Note: The graphs show estimates and confidence intervals from estimating (1) by cohort. Outcome is an index summarizing the sub-indices from the student survey presented in Figures A.14 and A.15. The different series correspond to treatment effects, δ in eq. (1), and placebo effects from the non-treatment margins, γ in eq. (1). The treated cohorts are shaded. The dashed vertical lines delimit the cohorts in their third year of treatment when they answer the survey. The figure notes show the sample mean and standard deviation of the outcome, estimated effect and standard errors for a pooled analysis of the treatment years, and the p-value of a joint test of all placebo effects (γ) for all treatment years. The regression uses school-level data with student weights.

Since we do not find any effects on academic outcomes, a skeptic might ask whether the school environment effects are just random artifacts of testing several outcomes and specifications. A multiple testing adjustment alleviates such concerns. Although we did not file a pre-analysis plan with a pre-determined number of tests, the paper is based on a policy evaluation commissioned by the Norwegian Directorate for Education and Training. Their call for proposals and the funded protocol entail some pre-registered decisions. Consistent with the overall aim of the reform (see Section 2) and the call, our funded protocol specified the outcomes of interest as school resources (with special needs teaching being an outcome and student-teacher ratios essentially a first stage), learning outcomes (including 9th grade test scores and exam scores), school environment, and medium-term outcomes (early measures of upper secondary school completion). Thus, the school environment is one out of four pre-specified domains and five pre-specified outcome measures. Our design also involves a choice of whether to include partly treated cohorts or not when averaging the effects for the longer-term outcomes, which was not specified in the proposal. Still, the number of main hypotheses tested is arguably smaller than or equal to 10. The p-values of the average effects on the school environment index are 0.018 for the 2013 and 2014 cohorts (fully treated) average and 0.001 for all the 2011-2016 cohorts. The latter is thus significant at the 5% level even with Bonferroni-adjustment for testing 10 hypotheses, which is known to be conservative, especially with non-independent tests. The average effects for the 2013-2014 cohorts is not, but this is due to lower precision, as the estimated effect on the school environment is largest for the fully treated. Thus, the school environment effects are robust to multiple testing adjustment and unlikely to be spurious.[15]

In light of the effects on the school environment, why are academic outcomes unaffected? Most importantly, the effect of more funds for teachers on the school environment is modest. To illustrate, an effect on the school environment of 0.05 SD (Figure 7) implies that one in twenty students answers that "most teachers" are supportive rather than just "a few teachers", or that one in twenty "enjoy school very much" rather than "enjoy school". Funding effects on academic outcomes that operate via a the school environment obviously depend on the impact of a slight improvement in the school environment on such outcomes. The association between the school environment index and student outcomes is shown in Appendix Table A.2. Unsurprisingly, cohorts of students that report a better environment also have more favorable outcomes, even within school conditional on student characteristics. While these associations may not be credible causal estimates,[16] they provide a relevant baseline for comparing estimated effects on the school environment and student outcomes.

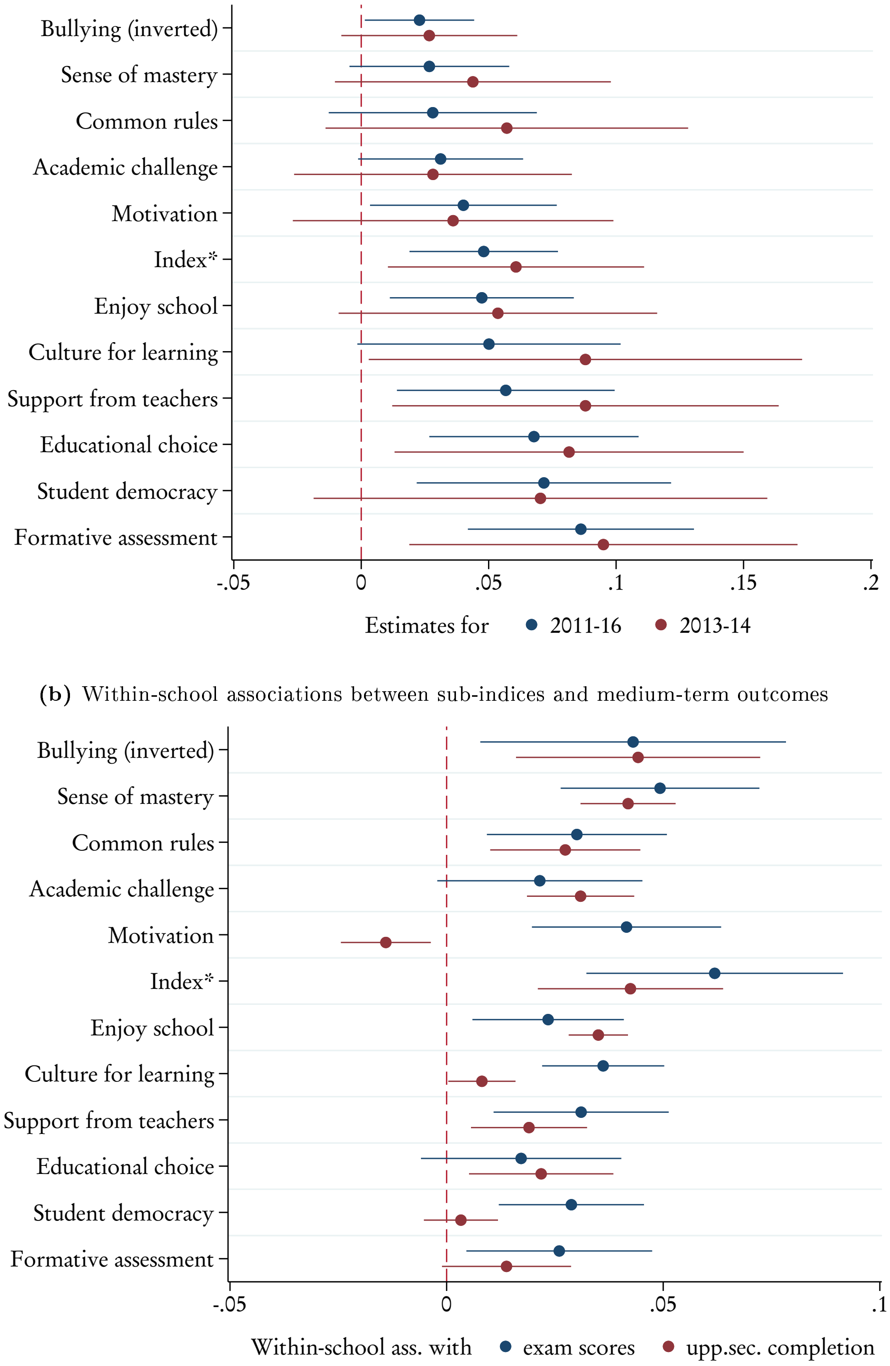
Even if a better school environment predicts improved educational outcomes, the within-school associations are also modest as shown in column (4) of Appendix Table A.2. If we multiply these effects of a better school environment on academic outcomes by the 5% improvement in the school environment (Figure 6), the funding of extra teachers is expected to improve students' exam scores by only 0.3% of a standard deviation and completion of the second year of upper secondary school by 0.2 percentage points as a result of the better school environment.

To better understand the divergent results for academic outcomes and the school environment, we also look for any pattern across aspects of the school environment. In Panel (a) of Figure 7, we summarize the treatment effects on the pooled cohorts both for the index from Figure 6 and for the separate sub-indices presented in Appendix Figures A.14 and A.15. All outcomes are measured in terms of comparable student-level standard deviations. Although all estimates are positive, not all are statistically significant, and the magnitude varies across dimensions. The effects are strongest for assessments that support learning (formative assessment), student democracy, guidance on educational choices, support from teachers, culture for learning, and school well-being, with smaller effects for bullying, sense of mastery, common rules, academic challenge, and school motivation. In all, we interpret the school environment results as improved student well-being. However, the funding appears to have the strongest impact on the school environment indicators that matter least for student outcomes. While Panel (a) of Figure 7 shows the effects of the extra teachers on the different sub-indices, Panel (b) shows within-school associations with student outcomes by sub-index. With the exception of motivation and completion of upper secondary school, all associations are positive, and mostly significantly so. However, while neither the effect estimates nor the within-school associations are sufficiently precise for equality to be rejected across most sub-indices, there is a clear negative correlation between the two. The estimated effects of extra

[15] The significant effect on the school environment is also robust with respect to the choice of global vs local specification, see Table A.8).

[16] There may still be unmeasured student characteristics contributing to the school environment and later outcomes. Also, by looking at within-school associations, we disregard how time-invariant school quality may impact both student outcomes and the school environment.

Figure 7. Effects on separate sub-indices and associations between sub-indices and medium-term outcomes



(a) Funding effects on separate sub-indices

(b) Within-school associations between sub-indices and medium-term outcomes

Note: Sub-graph (a) shows pooled effect estimates and confidence intervals of δ resulting from estimating (1) for 2011-2016 and 2013-2014 students, respectively. Sub-graph (b) shows point estimates and confidence intervals for within-school associations between the school environment indices and medium-term outcomes. Upper secondary school completion is measured five years after completing compulsory schooling. The index is an average of all the other sub-indices' outcomes. Outcomes are sorted by the 2011-2016 effect in sub-graph (a). All regressions use school-level data with student weights.

Table 2. Pooled treatment effect and placebo estimates

	Outcome					
	(1)	(2)	(3)	(4)	(5)	(6)
		9th grade	Exam	Teacher	Complete	School
	STR	test score	score	grades	year 2	env. index
Fully treated	2013-2016		2013-2014			
Treated	-2.293**	0.006	0.005	-0.033	-0.003	0.061**
	(0.250)	(0.011)	(0.025)	(0.030)	(0.009)	(0.026)
Placebo STR	0.320	0.005	0.006	0.057**	-0.001	0.011
	(0.212)	(0.010)	(0.021)	(0.026)	(0.007)	(0.022)
Placebo GPA	-0.020	-0.000	-0.027	-0.025	-0.000	-0.007
	(0.212)	(0.009)	(0.023)	(0.024)	(0.008)	(0.021)
N observations	3089	197,804	97,626	99,275	101,719	1438
Full and partly treated			2011-2016			
Treated			-0.004	-0.029	-0.002	0.048**
			(0.020)	(0.026)	(0.007)	(0.015)
Placebo STR			0.013	0.053**	-0.000	0.001
			(0.017)	(0.022)	(0.006)	(0.013)
Placebo GPA			-0.009	-0.012	0.002	-0.007
			(0.017)	(0.021)	(0.006)	(0.012)
N observations			291,389	295,907	254,372	4303

Note: The table reports the pooled effect estimates from Figures 3, 5, and 6, as well as corresponding pooled effects for the placebo margins. The analyses in columns (1) and (6) are at the school level, weighted with the number of students and with heteroskedasticity-robust standard errors. The analyses in columns (2)-(5) are at the student level, standard errors are adjusted for clustering at the school level. All analyses control for year and quadratics in the forcing variables (incl. interaction), the analyses in columns (2)-(5) also include controls for student characteristics (cubic in pretest, both parents' education, sex). ** p<.05, * p<.10.

teachers are greater on sub-indices that are more weakly related to educational outcomes.

7 Discussion

The intervention represented a significant investment of public funds that reduced the student-teacher ratio by about 10% in the school cohorts affected. More teachers led to an improved school environment as experienced by the students, but nothing happened to their academic outcomes. In Table 2 we summarize the main pooled effect estimates and the corresponding estimates from the placebo margins from the preceding sections. In this section, we offer possible explanations for why it did not improve academic outcomes.

7.1 Magnitudes and cost effectiveness

Standardized test scores, exam scores, and teacher-assessed grades have comparable scales since all are measured in terms of student-level standard deviations. The estimated effects on test scores are the most precisely estimated, allowing us to rule out effects larger than 3% of a standard deviation. These effects are estimated after just over one year of exposure and can be compared with those from the STAR project. While Krueger (1999) reports an effect size of 20% of a standard deviation (p.514), the reduction in class size was three times larger (30% compared to 10% in our case). Therefore, a 6-7% standard deviation increase is what to expect from an intervention like ours based on the STAR estimates, which is considerably larger than our upper bound of 3%.

Cost-effectiveness is a key test for any intervention. A zero-effect-intervention can never be efficient. Nevertheless, for policy-making it is also useful to consider cost-effectiveness of the upper-bound estimate. First, recall that the per-student costs in our intervention amount to USD 1,400 per year. To put a value on the potential effects on learning, we use the estimated association between school value-added and future earnings from Kirkebøen (2021), who finds that one standard deviation higher exam score increases future labor earnings in Norway by 1.5%. Using 2014 data, we estimate the average present value of labor earnings in Norway at age 13 to be USD 1.03 million.[17] Thus, the upper bound of the effect on exam scores (5.5% of an SD) implies an increase in discounted income of USD 854 per student, or 61% of the cost, while the value of an effect corresponding to the point estimate (0.5% of an SD) is only 6% of the cost. Even for the upper-bound estimate, the intervention is ineffective and we can rule out that more funds to lower the student-teacher ratio can be justified in terms of higher lifetime earnings for the students.

7.2 Possible explanations for the limited effects on academic outcomes

Teacher funding effects may depend on the institutional context, which raises the question of whether our results are policy-relevant outside of Norway. In particular, if there are diminishing marginal returns to school spending, we may expect additional funding to have small effects in a country like Norway, where spending is already high (Schleicher, 2018). Indeed, although the ratio of students to teaching staff is similar to many other European countries, OECD statistics show that it is slightly below the EU23 mean and the US mean (OECD, 2021).[18] However, this diminishing returns explanation is challenged by the fact that we identify marginal effects for students in schools with an above-average student-teacher ratio and below-average results. Moreover, there is no consensus on diminishing marginal returns to school spending; for example, Jackson and Mackevicius (2021) find little evidence of this in the US.

While we cannot explain why the extra funding to hire teachers did not improve academic outcomes, or how the intervention could have been designed to raise student performance, the data allow us to investigate some potential mechanisms. If school and parental inputs are substitutes in educational production, parents may reduce the effort and time devoted to their children's learning activities in response to a reduction in the student-teacher ratio. The evidence on parental reactions to changes in school inputs is far from conclusive, however (Rabe, 2019). For example, while a Swedish study points to lower parental effort when classes are small (Fredriksson et al.,

[17] This estimate is based on average earnings by age for the entire Norwegian population, discounted to age 13 of the student, i.e. at the start of lower secondary school, and summed over age: $\sum_{age=16}^{67}(1+discount)^{-(age-13)}\bar{y}_{age}$. The real discount rate is 4%, in line with what the Norwegian Ministry of Finance recommends for public investment.

[18] The ratio is 9 in Norway, compared to 11 in the EU23. The average class size is 23 for the EU23, not far from the ratio between regular teacher hours and student hours in this paper. Unfortunately, OECD does not include class size information for Norway.

2016), the opposite has been found in Norway (Bonesrønning, 2004). To test the potential part played by parental adjustment, we use a question in the student survey. Students are asked about to the extent to which parents show interest in school, help with homework, encourage them, and expect them to do their best. Using the same model as for the school environment effect(s), we find no indication that the presence of more teachers reduces parental support (Appendix Figure A.16). If anything, the (insignificant) positive coefficient suggests a complementary response rather than a substitution effect, as students perceive that parents increase their support when the student-teacher ratio is reduced. Thus, we find nothing indicating that the lack of effects on academic outcomes is explained by a reduction in parental input that counters an effect due to more teachers.

School principals and municipalities may potentially reallocate resources within or between schools as a response to the extra funding (see discussion in Hoxby, 2000). Since the effect on total teachers corresponds to the intended increase, there is no indication of between-school reallocation. Within-school reallocation can happen along several dimensions, studied in Appendix Figure A.17.[19] The funding increased regular teacher hours very similarly to total hours (including for example special needs education), and we find no change in teacher hours used for special needs teaching or special services for Norwegian language learners. Another possible margin of adjustment is teaching assistant hours, typically used to support special needs teaching. We do not find any effect on the ratio of assistant hours to student hours, nor any effect on assistant hours used for regular teaching.

Furthermore, if the treatment schools had difficulties recruiting competent teachers, we would expect to see reduced teacher qualifications in these schools (Gilraine, 2020). However, the increase in teacher hours taught by qualified teachers is very similar to the total effect on teacher hours. Additionally, there is no effect of the funding on the number of hours taught by teachers without formal qualifications (Appendix Figure A.18). Moreover, from matched employer-employee data (Appendix Figure A.19),[20] we find neither statistically significant nor quantitatively substantial effects on average time since teachers completed their education, tenure at the school, or average sickness absence of the teachers (which may be a measure both of (un)available teaching resources and of teacher workload, if a heavy workload induces illness and absences). However, there is a small increase in the share of teachers with a teaching degree in treatment schools.

The fact that we find few effects on teacher characteristics despite a substantial relative increase in the number of teachers reflects characteristics of the reform and the teacher labor market. While the extra funding substantially increased the number of teachers in the treated schools, these schools represent only about a quarter of total lower secondary school students (and a smaller share of schools and teachers, as the treatment schools are large schools with high student-teacher ratios). Furthermore, teachers may move between primary, lower secondary, and upper secondary schools. Also, there is substantial mobility in the teacher labor market, such that even in the treatment schools, new hires only increased moderately in response to the reform. Overall, while crowding out of other school input and recruitment constraints are potential explanations for the null effects on student academic outcomes, the evidence provides no basis for this interpretation.

[19] In Figures A.17 and A.18, we study teacher-student ratios, rather than student-teacher ratios, in order to have a fixed denominator and be able to decompose the teacher hours in the nominator. The denominator is total student hours, including special needs teaching, and thus slightly different from the nominator of our measure of student-teacher ratio. However, this difference is minor for student hours.

[20] This data source is available only for 2015-2018 cohorts.

When interpreting our null findings of teacher funding on academic outcomes, it is important to note that the schools were free to use the resources they deemed best. The effects of teacher funding may depend on the specific use of the extra teacher resources, and we cannot rule out that more targeted use of the resources may have improved academic outcomes. This interpretation is supported by recent evidence using RCT designs in Norway that has found that small-group instruction in mathematics indeed improves national test scores (Bonesrønning et al., 2022; Kirkebøen et al., 2021). In supplementary analyses, we explore whether funding effects depend on whether the schools use small-group instruction (Appendix Figure A.20), split students by academic ability (Appendix Figure A.21), prioritize the lowest-performing students (Appendix Figure A.22), and use extra teachers in most 8th-grade mathematics lessons (Appendix Figure A.23). Across all dimensions, there are few significant single-year effect estimates, and treatment effects are similar for different usage of the extra teacher resouces. However, these results do not provide strong evidence on the impact of specific use of teacher resources. For example, among schools using small-group instructions, we do not know whether such resources were used consistently over an extended period, nor which subjects it was used in. Furthermore, estimating heterogeneous effects by characteristics of the implementation is methodologically challenging.[21] Thus, our study mainly speaks to the effects of hiring extra teachers irrespective of how these teachers were used.

8 Conclusion

We have studied a large-scale intervention that funded 600 extra teachers per year distributed among 166 lower secondary schools in Norway for a four-year period, comparable to an increase in cost per student of USD 1,400 in each of the three years of lower secondary education. The funding assignment was based on two sharp conditions, allowing us to credibly identify the effects of additional resources using a regression discontinuity framework. We find that the extra funding reduced the student-teacher ratio by around 10%. There is no evidence of crowding out of other school inputs and the treatment schools did not seem to meet any restrictions when recruiting additional teachers. Nor did the policy induce other compensatory adjustments, such as changes in teacher composition, teacher sickness absence, or special needs education. The reduced student-teacher ratio did not improve academic outcomes, as measured by 9th grade test scores, end-of compulsory schooling grades, or upper secondary school progression. Inspired by recent evidence of the effects of school inputs on non-cognitive outcomes, we also tested whether more teachers changed how students perceived the school environment. Although the funding improved several aspects of the school environment, including student well-being, these effects were too small to significantly impact academic outcomes.

A major advantage of this study is that its findings have clear policy implications. In the quest for causal identification, studies often need to exploit natural experiments at margins far

[21] While we can study outcomes in treated schools that used resources differently, we do not know which control schools would have implemented the intervention in a given way and, therefore, which control schools constitute a valid control group for each treatment. Nevertheless, with our data, the only feasible approach is to estimate treatment effects for each treatment compared to all control schools. We use the (placebo) effect estimates from pre-treatment years to investigate whether there are time-invariant differences between treatment schools implementing different treatments. There are no pre-treatment differences, suggesting that the funding effects for schools using resources in different ways could be interpreted causally.

from relevant policy changes. In our study, however, credible identification and policy relevance go hand in hand. We directly identify the parameter of interest for funding policies set by the national government: Does increasing funding to provide more teachers improve student outcomes? For countries like Norway, our evidence shows that when schools are free to use the additional teachers in grade 8-10 as they see fit, the school environment improves, but the impacts on students' academic achievement and upper secondary school dropout are negligible.

References

Abott, C., V. Kogan, S. Lavertu, and Z. Peskowitz (2020). School district operational spending and student outcomes: Evidence from tax elections in seven states. *Journal of Public Economics 183*, 104–142.

Angrist, J. D., E. Battistin, and D. Vuri (2017). In a small moment: Class size and moral hazard in the italian mezzogiorno. *American Economic Journal: Applied Economics 9*(4), 216–49.

Angrist, J. D. and V. Lavy (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics 114*(2), 533–575.

Angrist, J. D., V. Lavy, J. Leder-Luis, and A. Shany (2019). Maimonides' rule redux. *American Economic Review: Insights 1*(3), 309–24.

Argaw, B. A. and P. A. Puhani (2018). Does class size matter for school tracking outcomes after elementary school? quasi-experimental evidence using administrative panel data from germany. *Economics of Education Review 65*, 48–57.

Baron, E. J. (2022). School spending and student outcomes: Evidence from revenue limit elections in wisconsin. *American Economic Journal: Economic Policy*.

Bonesrønning, H. (2003). Class size effects on student achievement in norway: Patterns and explanations. *Southern Economic Journal*, 952–965.

Bonesrønning, H. (2004). The determinants of parental effort in education production: do parents respond to changes in class size? *Economics of Education Review 23*(1), 1–9.

Bonesrønning, H., H. Finseraas, I. Hardoy, J. M. V. Iversen, O. H. Nyhus, V. Opheim, K. V. Salvanes, A. M. J. Sandsør, and P. Schøne (2022). Small-group instruction to improve student performance in mathematics in early grades: Results from a randomized field experiment. *Journal of Public Economics 216*, 104765.

Browning, M. and E. Heinesen (2007). Class size, teacher hours and educational attainment. *Scandinavian Journal of Economics 109*(2), 415–438.

Brunner, E., B. Hoen, and J. Hyman (2022). School district revenue shocks, resource allocations, and student achievement: Evidence from the universe of us wind energy installations. *Journal of Public Economics 206*, 104586.

Brunner, E., J. Hyman, and A. Ju (2020). School finance reforms, teachers' unions, and the allocation of school resources. *Review of Economics and Statistics 102*(3), 473–489.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust data-driven inference in the regression-discontinuity design. *The Stata Journal 14*(4), 909–946.

Cattaneo, M. D., R. Titiunik, and G. Vazquez-Bare (2020). Analysis of regression-discontinuity designs with multiple cutoffs or multiple scores. *The Stata Journal 20*(4), 866–891.

Chay, K. Y., P. J. McEwan, and M. Urquiola (2005). The central role of noise in evaluating interventions that use test scores to rank schools. *The American Economic Review 95*(4), 1237–1258.

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

Chingos, M. M. (2012). The impact of a universal class-size reduction policy: Evidence from florida's statewide mandate. *Economics of Education Review 31*(5), 543–562.

Cornelissen, T. and C. Dustmann (2019). Early school exposure, test scores, and noncognitive outcomes. *American Economic Journal: Economic Policy 11*(2), 35–63.

Dee, T. S. and M. R. West (2011). The non-cognitive returns to class size. *Educational Evaluation and Policy Analysis 33*(1), 23–46.

Deming, D. J. (2022). Four facts about human capital. *Journal of Economic Perspectives 36*(3), 75–102.

Falch, T., A. M. J. Sandsør, and B. Strøm (2017). Do smaller classes always improve students' long-run outcomes? *Oxford Bulletin of Economics and Statistics 79*(5), 654–688.

Fredriksson, P., B. Öckert, and H. Oosterbeek (2013). Long-term effects of class size. *The Quarterly Journal of Economics 128*(1), 249–285.

Fredriksson, P., B. Öckert, and H. Oosterbeek (2016). Parental responses to public investments in children: Evidence from a maximum class size rule. *Journal of Human Resources 51*(4), 832–868.

Gibbons, S., S. McNally, and M. Viarengo (2018). Does additional spending help urban schools? an evaluation using boundary discontinuities. *Journal of the European Economic Association 16*(5), 1618–1668.

Gilraine, M. (2020). A method for disentangling multiple treatments from a regression discontinuity design. *Journal of Labor Economics 38*(4), 1267–1311.

Hægeland, T., O. Raaum, and K. G. Salvanes (2012). Pennies from heaven? using exogenous tax variation to identify effects of school resources on pupil achievement. *Economics of Education Review 31*(5), 601–614.

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

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics 115*(4), 1239–1285.

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy 126*(5), 2072–2107.

Jackson, C. K. (2020). Does school spending matter? the new literature on an old question. In R. D. L. Tach and D. L. Miller (Eds.), *Confronting inequality: How policies and practices shape children's opportunities*. American Psychological Association.

Jackson, C. K. and C. Mackevicius (2021). The distribution of school spending impacts. *NBER WP* (Number 28517).

Jackson, C. K., S. C. Porter, J. Q. Easton, A. Blanchard, and S. Kiguel (2020). School effects on socioemotional development, school-based arrests, and educational attainment. *American Economic Review: Insights 2*(4), 491–508.

Jepsen, C. and S. Rivkin (2009). Class size reduction and student achievement: The potential tradeoff between teacher quality and class size. *Journal of Human Resources 44*(1), 223–250.

Kirkebøen, L. (2021). School value-added and long-term student outcomes. Discussion Papers 970, Statistics Norway, Research Department.

Kirkebøen, L. J., T. Gunnes, L. Lindenskov, and M. Rønning (2021). Didactic methods and small-group instruction for low-performing adolescents in mathematics. results from a randomized controlled trial. Discussion Papers 957, Statistics Norway, Research Department.

Kirkebøen, L. J., A. Kotsadam, O. Raaum, S. Andresen, and J. Rogstad (2017). Effekter av satsing på økt lærertetthet. *Mimeo SSB*.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics 114*(2), 497–532.

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

Leuven, E. and S. A. Løkken (2020). Long-term impacts of class size in compulsory school. *Journal of Human Resources 55*(1), 309–348.

Leuven, E., H. Oosterbeek, and M. Rønning (2008). Quasi-experimental estimates of the effect of class size on achievement in norway. *Scandinavian Journal of Economics 110*(4), 663–693.

OECD (2021). *Education at a Glance 2021. OECD Indicators.* OECD Publishing.

Rabe, B. (2019). Do school inputs crowd out parents investments in their children? *IZA World of Labor*.

Reiling, R. B., K. V. Salvanes, A. M. J. Sandsør, and B. Strøm (2021). The effect of central government grants on local educational policy. *European Journal of Political Economy 69*, 102006.

Schleicher, A. (2018). Insights and interpretations. *Pisa 2018 10*.

Todd, P. E. and K. I. Wolpin (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal 113*(485), F3–F33.

Urquiola, M. (2006). Identifying class size effects in developing countries: Evidence from rural bolivia. *Review of Economics and Statistics 88*(1), 171–177.

Woessmann, L. and M. West (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in timss. *European Economic Review 50*(3), 695–736.

A Results referred to in the text

Table A.1. Compensating school inputs. Association between entry test score and student teacher ratio in the pre-treatment school years (2011-2012)

	All schools		Large municipalities
	(1)	(2)	(3)
A. Student-teacher ratio			
Entry test score (8th grade)	3.490**	2.430**	3.235**
	(0.309)	(0.357)	(0.527)
Mean (std dev)	15.9 (2.8)		17.9 (2.2)
B. Student-teacher ratio in regular teaching			
Entry test score (8th grade)	2.568**	0.884**	1.461**
	(0.364)	(0.450)	(0.620)
Mean (std dev)	20.0 (3.2)		22.0 (2.5)
Fixed effects	Year	Year*Municipality	Year*Municipality
N schools	1 574	1 574	239
N students	336 556	336 556	75 665

Note: Each cell is an estimate from a separate model that regresses resource inputs in grades 8-10 on average entry test score in 8th grade (beginning of lower secondary) for different samples and specifications. Sample of pre-treatment school years 2011-2012. For a year t, test scores are the average 8th grade scores for the years $t, t-1, t-2$. The top row shows total student hours/total teacher hours (Panel A) and the bottom row shows regular student instruction hours/regular teaching hours (Panel B). Column (1) controls for year fixed effects, columns (2) and (3) control for Year*Municipality fixed effects. In column (3) we restrict the sample to large municipalities (defined as having more than 9 schools). Standard errors clustered at school level in parentheses. ** p<.05.

Table A.2. Associations between school environment index and academic outcomes

	(1)	(2)	(3)	(4)
Teacher-assigned grades (10th grade)	0.319**	0.069**	0.235**	0.098**
	(0.027)	(0.019)	(0.014)	(0.013)
Exam scores (10th grade)	0.349**	0.127**	0.180**	0.062**
	(0.030)	(0.017)	(0.017)	(0.015)
Completion of second year of upper secondary school	0.075**	0.045**	0.076**	0.058**
	(0.007)	(0.006)	(0.006)	(0.006)
Completion of upper secondary (within 5 years)	0.088**	0.043**	0.066**	0.042**
	(0.009)	(0.007)	(0.012)	(0.011)
Control variables	No	Yes	No	Yes
Fixed effects	No	No	Yes	Yes

Note: Each cell is an estimate from a separate model that regresses academic outcomes on the school environment index among cohorts from 2011 to the latest year for which figures are available (2016 for exam score, 2012 for completion of upper secondary). Standard errors clustered at school level in parentheses. ** p<.05.

Table A.3. Robustness - regular-teaching student teacher ratio

	(1 - baseline)	(2)	(3)	(4)
Treatment	-2.293**	-2.779**	-2.105**	-2.348**
	(0.250)	(0.196)	(0.345)	(0.260)
STR margin	0.320	0.533**	0.062	0.236
	(0.212)	(0.195)	(0.263)	(0.232)
GPA margin	-0.020	0.192	0.019	0.102
	(0.212)	(0.198)	(0.253)	(0.229)
Specification	Global		Local	
Forcing variables	quadratic	linear	quadratic	linear
N	3089	3089	1687	1687

Note: Data are schools 2013-2016. Specification (1) corresponds to the main specification as in Figure 3 and Table 2. Specifications (1)-(2) use all schools. Specifications (3)-(4) use school within .5 SD of either threshold and triangular weights (similar to the RD analyses). Specifications (2) and (4) use linear controls for the forcing variables, while (1) and (3) use quadratic controls. All analyses are at the school level, weighted with the number of students. ** p<.05, * p<.10.

Table A.4. Robustness - 9th grade test score

	(1 - baseline)	(2)	(3)	(4)	(5)	(6)
Treatment	0.006	0.046	-0.074**	0.019	0.097*	0.033
	(0.011)	(0.035)	(0.030)	(0.017)	(0.053)	(0.040)
STR margin	0.005	-0.002	0.067**	-0.002	-0.053	-0.013
	(0.010)	(0.033)	(0.031)	(0.013)	(0.041)	(0.038)
GPA margin	-0.000	0.011	0.070**	-0.019	-0.034	-0.007
	(0.009)	(0.028)	(0.029)	(0.012)	(0.036)	(0.034)
Specification	Global			Local		
Forcing variables	quadratic	quadratic	linear	quadratic	quadratic	linear
Background chars.	yes			yes		
N	197804	205130	205130	119343	123572	123572

Note: Data are students 2013-2016. Specification (1) corresponds to the main specification as in Figure 5 and Table 2. Specifications (1)-(3) use all schools. Specifications (4)-(6) use school within .5 SD of either threshold and triangular weights (similar to the RD analyses). Specifications (1), (2), (4) and (5) use quadratic controls for the forcing variables, while (3) and (6) use linear controls. Specifications (1) and (4) include controls for student characteristics (cubic in pretest, both parents' education, sex). All analyses are at the student level, standard errors are adjusted for clustering at the school level. ** p<.05, * p<.10.

Table A.5. Robustness - exam score

	(1 - baseline)	(2)	(3)	(4)	(5)	(6)
Treatment	0.005	0.026	-0.078**	-0.025	0.038	0.006
	(0.025)	(0.038)	(0.031)	(0.036)	(0.056)	(0.043)
STR margin	0.006	-0.003	0.056*	0.032	-0.015	0.007
	(0.021)	(0.035)	(0.032)	(0.028)	(0.043)	(0.040)
GPA margin	-0.027	0.007	0.057*	-0.010	0.000	0.012
	(0.023)	(0.033)	(0.031)	(0.028)	(0.041)	(0.038)
Specification	Global			Local		
Forcing variables	quadratic	quadratic	linear	quadratic	quadratic	linear
Background chars.	yes			yes		
N	97626	101032	101032	58922	60867	60867

Note: Data are students 2013-2014. Specification (1) corresponds to the main specification (fully treated) as in Figure 5 and Table 2. Specifications (1)-(3) use all schools. Specifications (4)-(6) use school within .5 SD of either threshold and triangular weights (similar to the RD analyses). Specifications (1), (2), (4) and (5) use quadratic controls for the forcing variables, while (3) and (6) use linear controls. Specifications (1) and (4) include controls for student characteristics (cubic in pretest, both parents' education, sex). All analyses are at the student level, standard errors are adjusted for clustering at the school level. ** p<.05, * p<.10.

Table A.6. Robustness - teacher grades

	(1 - baseline)	(2)	(3)	(4)	(5)	(6)
Treatment	-0.033	-0.006	-0.046*	-0.010	0.053	-0.000
	(0.030)	(0.032)	(0.025)	(0.046)	(0.050)	(0.035)
STR margin	0.057**	0.052*	0.076**	0.049	0.001	0.032
	(0.026)	(0.028)	(0.026)	(0.035)	(0.040)	(0.033)
GPA margin	-0.025	0.008	0.026	-0.013	-0.000	0.024
	(0.024)	(0.024)	(0.023)	(0.032)	(0.029)	(0.027)
Specification	Global			Local		
Forcing variables	quadratic	quadratic	linear	quadratic	quadratic	linear
Background chars.	yes			yes		
N	99275	102767	102767	59922	61922	61922

Note: Data are students 2013-2014. Specification (1) corresponds to the main specification (fully treated) as in Figure 5 and Table 2. Specifications (1)-(3) use all schools. Specifications (4)-(6) use school within .5 SD of either threshold and triangular weights (similar to the RD analyses). Specifications (1), (2), (4) and (5) use quadratic controls for the forcing variables, while (3) and (6) use linear controls. Specifications (1) and (4) include controls for student characteristics (cubic in pretest, both parents' education, sex). All analyses are at the student level, standard errors are adjusted for clustering at the school level. ** p<.05, * p<.10.

Table A.7. Robustness - on-time completion of second year in upper secondary

	(1 - baseline)	(2)	(3)	(4)	(5)	(6)
Treatment	-0.003	0.000	-0.010	0.003	0.008	-0.002
	(0.009)	(0.011)	(0.008)	(0.016)	(0.019)	(0.013)
STR margin	-0.001	0.001	0.006	-0.003	-0.004	0.001
	(0.007)	(0.009)	(0.008)	(0.011)	(0.014)	(0.011)
GPA margin	-0.000	0.001	0.006	-0.008	-0.009	-0.004
	(0.008)	(0.009)	(0.009)	(0.011)	(0.013)	(0.011)
Specification		Global			Local	
Forcing variables	quadratic	quadratic	linear	quadratic	quadratic	linear
Background chars.	yes			yes		
N	101719	106733	106733	61311	64246	64246

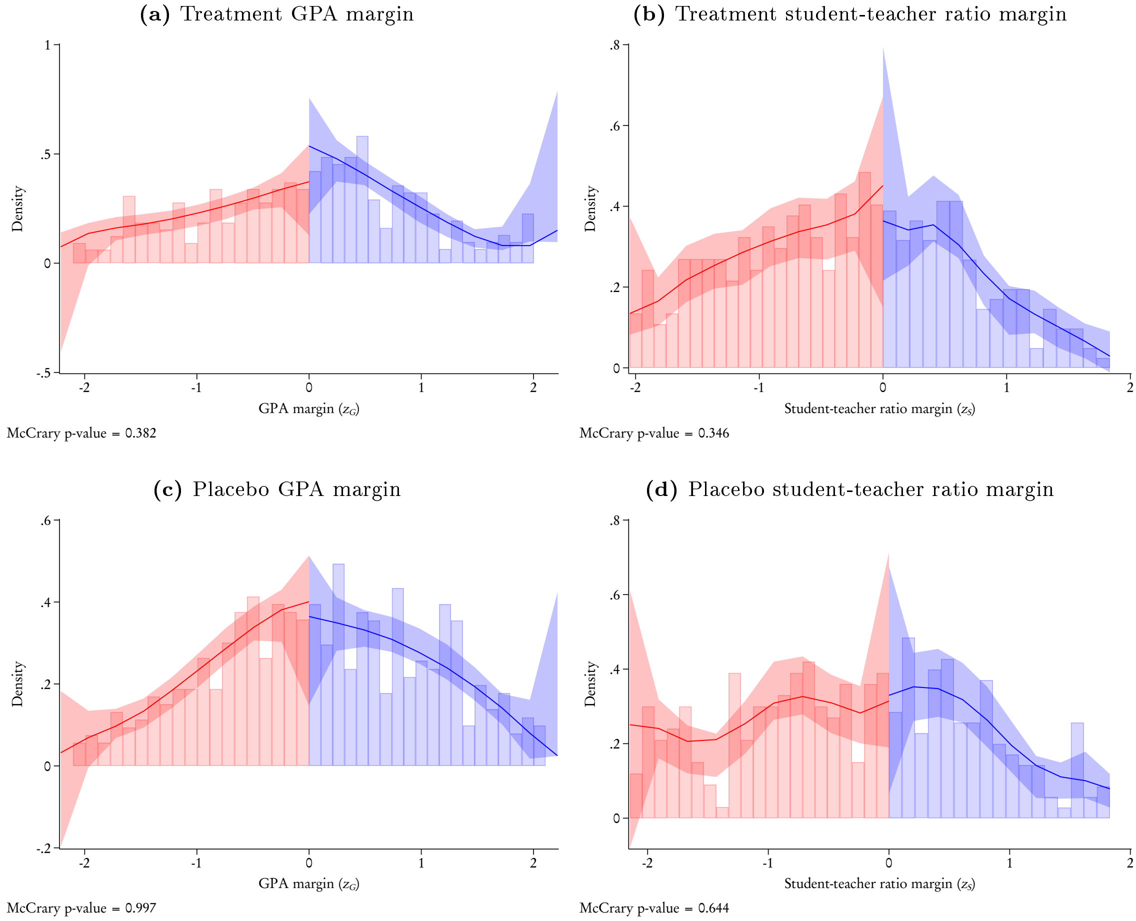
Note: Data are students 2013-2014. Specification (1) corresponds to the main specification (fully treated) as in Figure 5 and Table 2. Specifications (1)-(3) use all schools. Specifications (4)-(6) use school within .5 SD of either threshold and triangular weights (similar to the RD analyses). Specifications (1), (2), (4) and (5) use quadratic controls for the forcing variables, while (3) and (6) use linear controls. Specifications (1) and (4) include controls for student characteristics (cubic in pretest, both parents' education, sex). All analyses are at the student level, standard errors are adjusted for clustering at the school level. ** p<.05, * p<.10.

Table A.8. Robustness - school environment index

	(1 - baseline)	(2)	(3)	(4)
Treatment	0.061**	0.041**	0.112**	0.093**
	(0.024)	(0.019)	(0.033)	(0.024)
STR margin	0.011	0.022	-0.007	0.003
	(0.021)	(0.019)	(0.025)	(0.022)
GPA margin	-0.007	0.003	-0.027	-0.018
	(0.021)	(0.019)	(0.024)	(0.021)
Specification	Global		Local	
Forcing variables	quadratic	linear	quadratic	linear
N	1438	1438	802	802

Note: Data are schools 2013-2014. Specification (1) corresponds to the main specification as in Figure 6 and Table 2. Specifications (1)-(2) use all schools. Specifications (3)-(4) use school within .5 SD of either threshold and triangular weights (similar to the RD analyses). Specifications (2) and (4) use linear controls for the forcing variables, while (1) and (3) use quadratic controls. All analyses are at the school level, weighted with the number of students. ** p<.05, * p<.10.

Figure A.1. Balancing


(a) Treatment GPA margin

McCrary p-value = 0.382

(b) Treatment student-teacher ratio margin

McCrary p-value = 0.346

(c) Placebo GPA margin

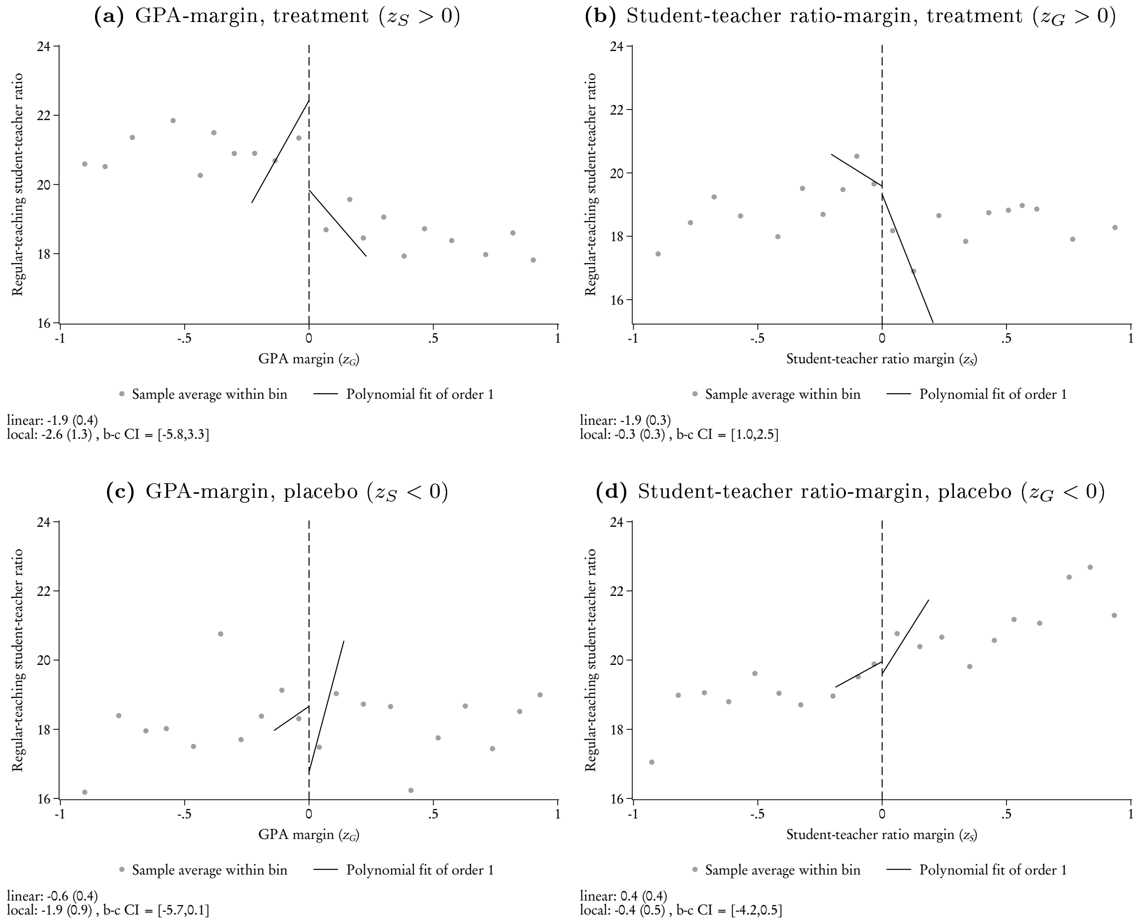
McCrary p-value = 0.997

(d) Placebo student-teacher ratio margin

McCrary p-value = 0.644

Note: The graphs show RD density estimates. Data are from the 2011, the base year for assigning the treatment. Figure notes show p-values for a test of discontinuity at the threshold. Densities and p-values are estimated using rddensity Calonico et al. (2014), using school-level data.

Figure A.2. RD-estimates of the effect on student-teacher ratio, RMSE-optimal bandwidths



(a) linear: -1.9 (0.4)
local: -2.6 (1.3) , b-c CI = [-5.8,3.3]

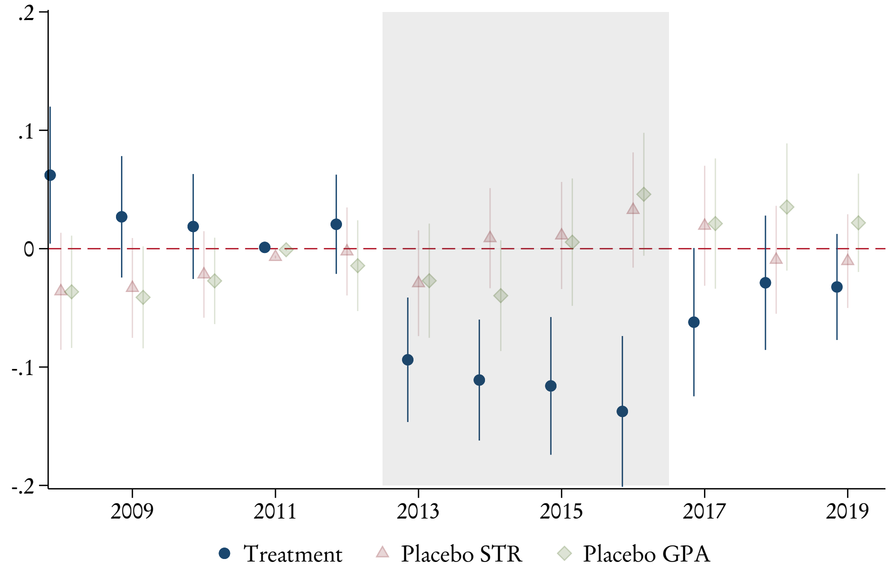
(b) linear: -1.9 (0.3)
local: -0.3 (0.3) , b-c CI = [1.0,2.5]

(c) linear: -0.6 (0.4)
local: -1.9 (0.9) , b-c CI = [-5.7,0.1]

(d) linear: 0.4 (0.4)
local: -0.4 (0.5) , b-c CI = [-4.2,0.5]

Note: The graphs show local RD-estimates at the different treatment and placebo margins. Data are years 2013-2016, outcome is (regular-teaching) student-teacher ratio. Figure notes show coefficients and standard errors from linear and local regressions. All analyses uses student weights. The lines show the local linear regressions and are estimated using Calonico et al. (2014), with triangular weights and RMSE-optimal bandwidth . Bias-corrected confidence interval (estimated using higher-order polynomial) in brackets. Bins are quantile-based.

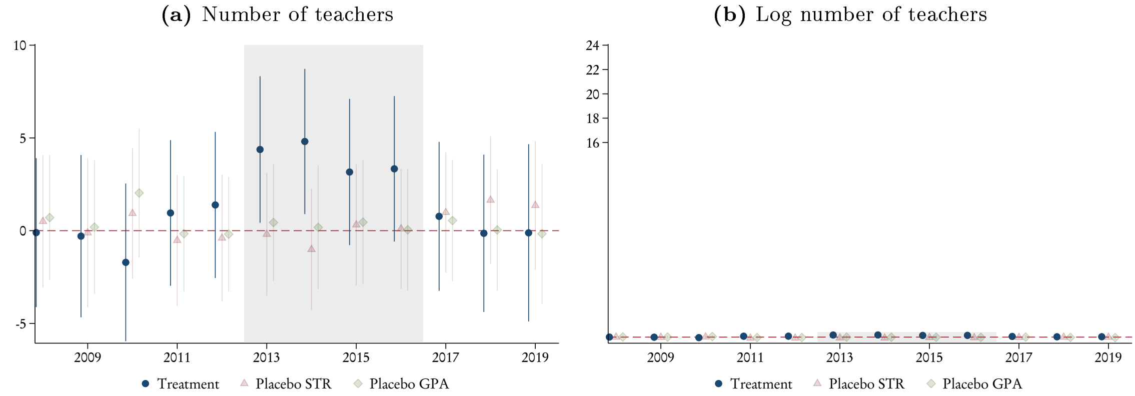
Figure A.3. Effects on log student-teacher ratio



sample mean (SD) = 2.866 (0.227)
pooled 2013-2016 estimate = -0.114** (0.014)
p-value joint test of 2013-2016 placebos = 0.2623

Note: The graph show estimates and confidence intervals from estimating (1). Outcome is log of (regular-teaching) student-teacher ratio. The regression uses student weights. See note to Figure 3 for further details.

Figure A.4. Effects on number of teachers


(a) Number of teachers

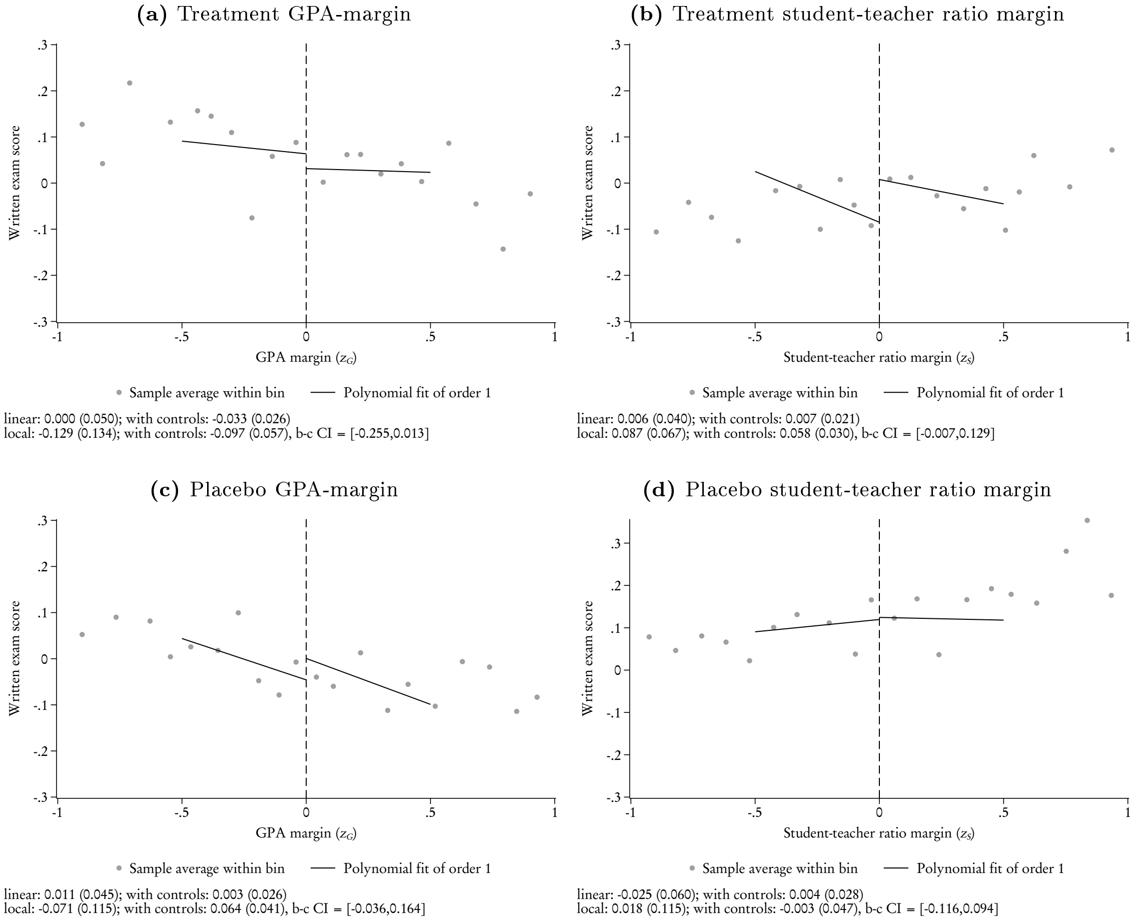
(b) Log number of teachers

sample mean (SD) = 21.1 (11.5)
pooled 2013-2016 estimate = 3.9** (1.0)
p-value joint test of 2013-2016 placebos = 0.9995

sample mean (SD) = 2.874 (0.642)
pooled 2013-2016 estimate = 0.159** (0.040)
p-value joint test of 2013-2016 placebos = 0.9338

Note: The graph shows estimates and confidence intervals from estimating (1). Outcomes are number of teacher at the school and log number of teachers. The regression uses student weights. See note to Figure 3 for further details.

Figure A.5. RD estimates of effects on exam score


(a) Treatment GPA-margin

(b) Treatment student-teacher ratio margin

(c) Placebo GPA-margin

(d) Placebo student-teacher ratio margin

(a) linear: 0.000 (0.050); with controls: -0.033 (0.026)
local: -0.129 (0.134); with controls: -0.097 (0.057), b-c CI = [-0.255,0.013]

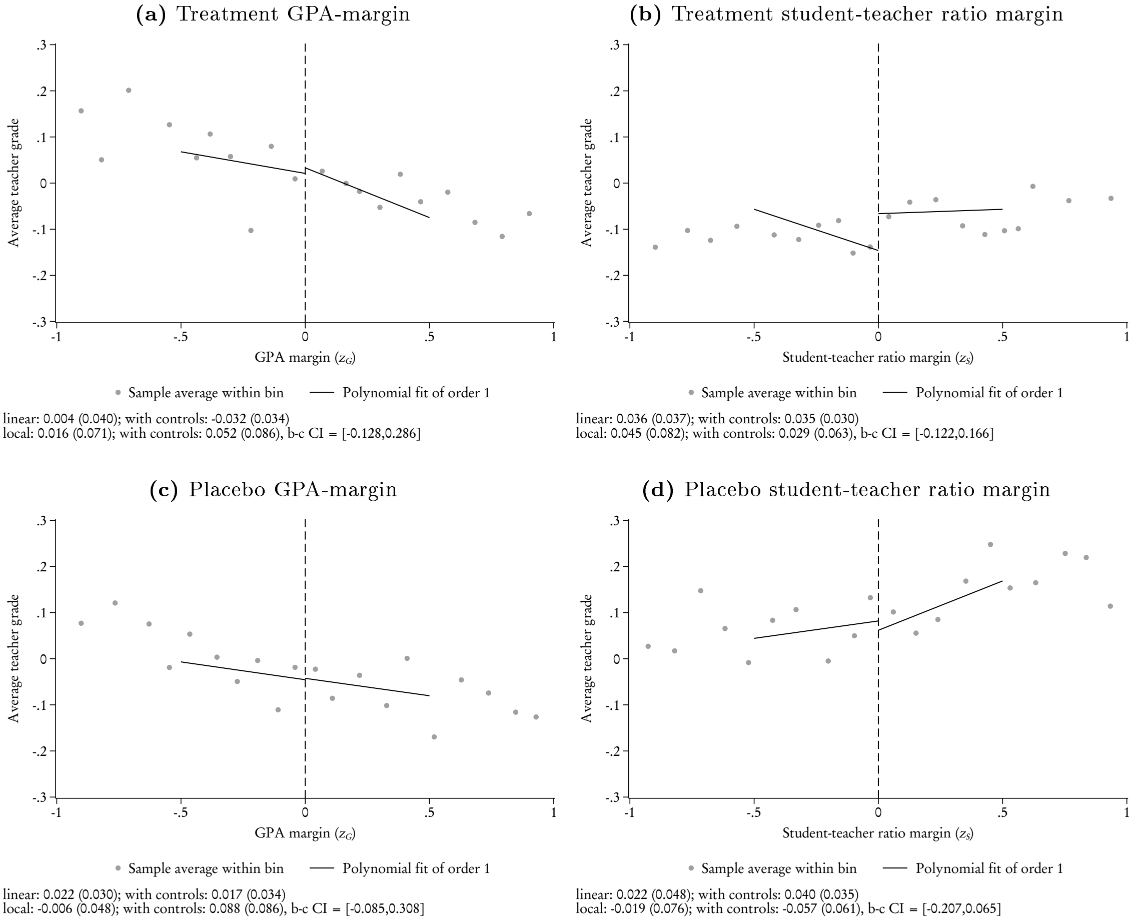
(b) linear: 0.006 (0.040); with controls: 0.007 (0.021)
local: 0.087 (0.067); with controls: 0.058 (0.030), b-c CI = [-0.007,0.129]

(c) linear: 0.011 (0.045); with controls: 0.003 (0.026)
local: -0.071 (0.115); with controls: 0.064 (0.041), b-c CI = [-0.036,0.164]

(d) linear: -0.025 (0.060); with controls: 0.004 (0.028)
local: 0.018 (0.115); with controls: -0.003 (0.047), b-c CI = [-0.116,0.094]

Note: The graphs show RD estimates for exam scores. The data are provided by students sitting the grade 9 test in 2013-2016. The graph is based on school-level data and student weights. The lines show the local linear regressions and are estimated using Calonico et al. (2014), with no additional control variables, triangular weights and a fixed bandwidth of 0.5. Bins are quantile-based. Figure notes show coefficients and standard errors from student-level linear and local regressions. Student controls in the linear and local regressions include gender, year dummies, a cubic in the 8th grade test score, and parental education. Bias-corrected confidence interval (estimated using higher-order polynomial) in brackets.

Figure A.6. RD estimates of effects on teacher grades



(a) Treatment GPA-margin

(b) Treatment student-teacher ratio margin

(c) Placebo GPA-margin

(d) Placebo student-teacher ratio margin

(a) linear: 0.004 (0.040); with controls: -0.032 (0.034)
local: 0.016 (0.071); with controls: 0.052 (0.086), b-c CI = [-0.128,0.286]

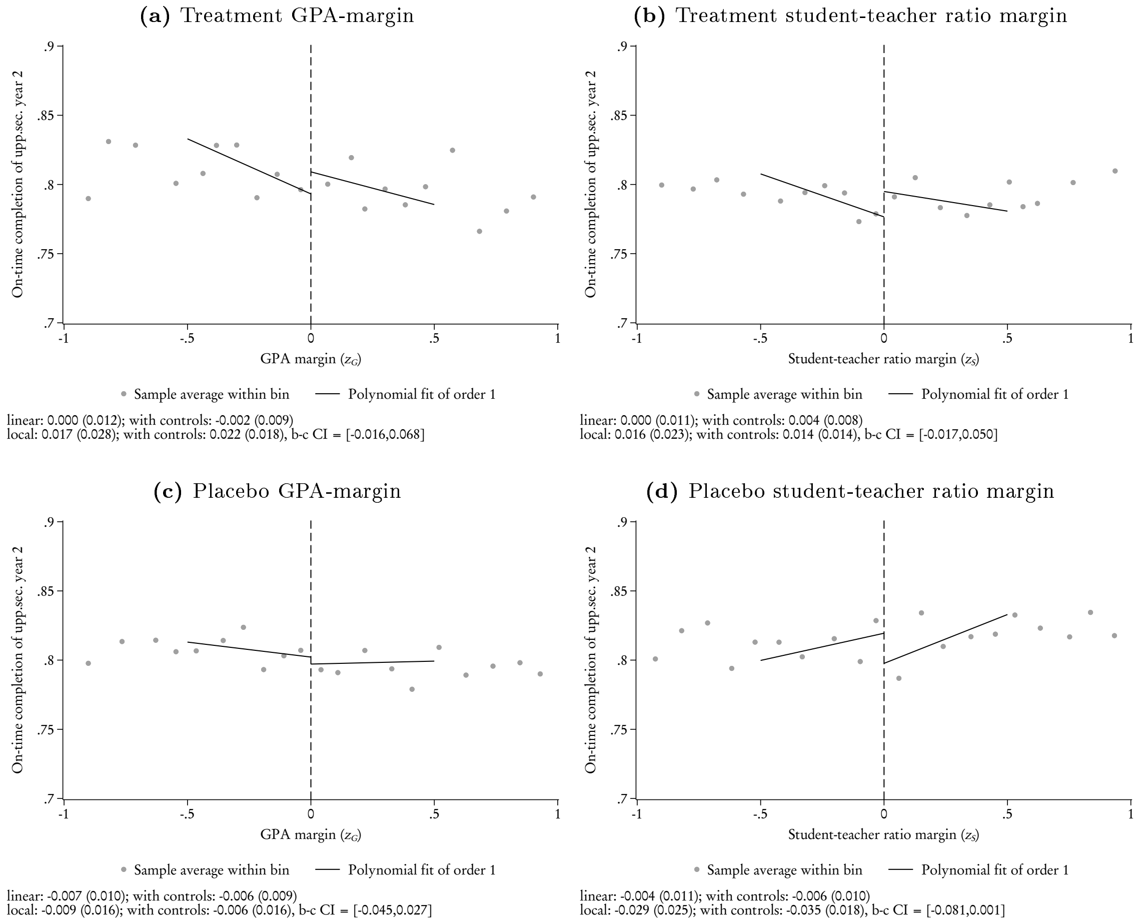
(b) linear: 0.036 (0.037); with controls: 0.035 (0.030)
local: 0.045 (0.082); with controls: 0.029 (0.063), b-c CI = [-0.122,0.166]

(c) linear: 0.022 (0.030); with controls: 0.017 (0.034)
local: -0.006 (0.048); with controls: 0.088 (0.086), b-c CI = [-0.085,0.308]

(d) linear: 0.022 (0.048); with controls: 0.040 (0.035)
local: -0.019 (0.076); with controls: -0.057 (0.061), b-c CI = [-0.207,0.065]

Note: The graphs show RD estimates for the average teacher grades. The data are provided by students sitting the grade 9 test in 2013-2016. The graph is based on school-level data and student weights. The lines show the local linear regressions and are estimated using Calonico et al. (2014), with no additional control variables, triangular weights and a fixed bandwidth of 0.5. Bins are quantile-based. Figure notes show coefficients and standard errors from student-level linear and local regressions. Student controls in the linear and local regressions include gender, year dummies, a cubic in the 8th grade test score, and parental education. Bias-corrected confidence interval (estimated using higher-order polynomial) in brackets.

Figure A.7. RD estimates of effects on on-time completion of year two upper secondary



(a) Treatment GPA-margin

linear: 0.000 (0.012); with controls: -0.002 (0.009)
local: 0.017 (0.028); with controls: 0.022 (0.018), b-c CI = [-0.016,0.068]

(b) Treatment student-teacher ratio margin

linear: 0.000 (0.011); with controls: 0.004 (0.008)
local: 0.016 (0.023); with controls: 0.014 (0.014), b-c CI = [-0.017,0.050]

(c) Placebo GPA-margin

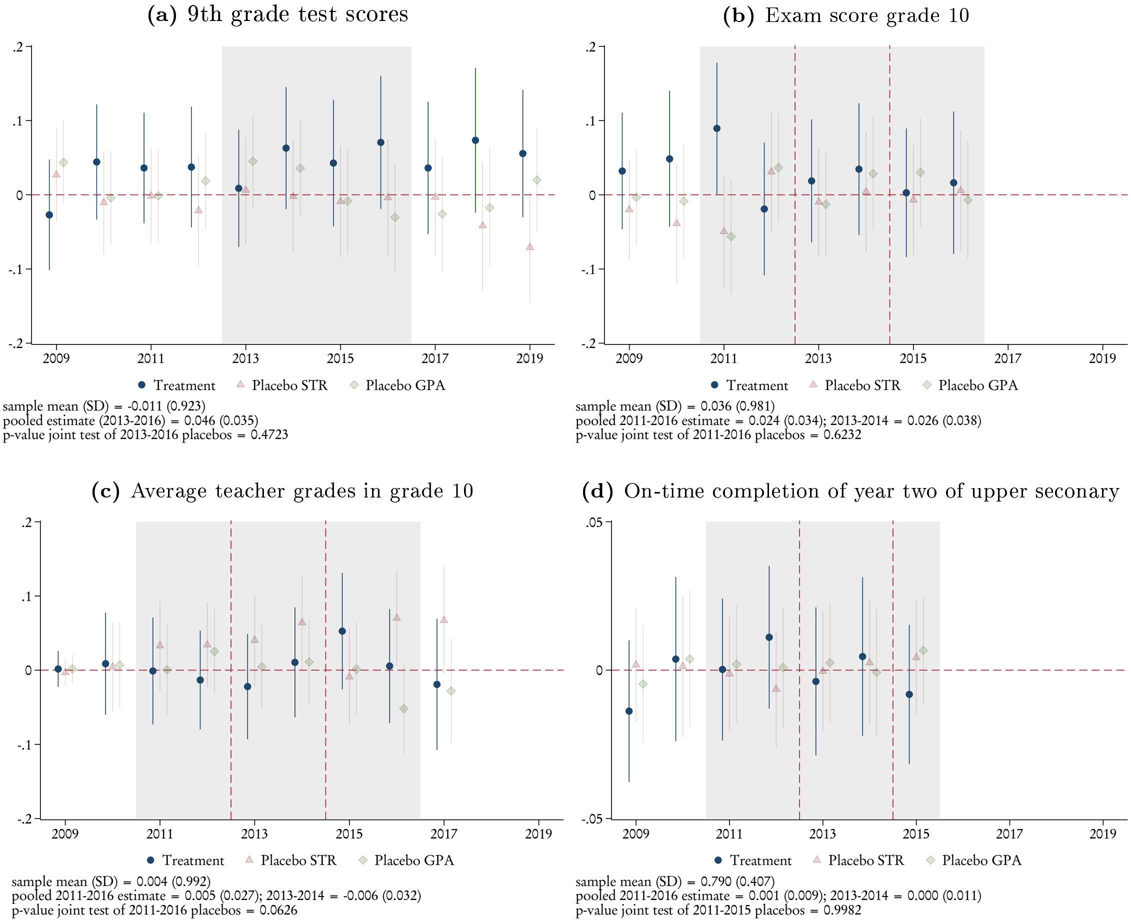
linear: -0.007 (0.010); with controls: -0.006 (0.009)
local: -0.009 (0.016); with controls: -0.006 (0.016), b-c CI = [-0.045,0.027]

(d) Placebo student-teacher ratio margin

linear: -0.004 (0.011); with controls: -0.006 (0.010)
local: -0.029 (0.025); with controls: -0.035 (0.018), b-c CI = [-0.081,0.001]

Note: The graphs show RD estimates for on-time completion of the second year of high school. The data are provided by students sitting the grade 9 test in 2013-2016. The graph is based on school-level data and student weights. The lines show the local linear regressions and are estimated using Calonico et al. (2014), with no additional control variables, triangular weights and a fixed bandwidth of 0.5. Bins are quantile-based. Figure notes show coefficients and standard errors from student-level linear and local regressions. Student controls in the linear and local regressions include gender, year dummies, a cubic in the 8th grade test score, and parental education. Bias-corrected confidence interval (estimated using higher-order polynomial) in brackets.

Figure A.8. Effects on academic outcomes, no student level controls



(a) 9th grade test scores

sample mean (SD) = -0.011 (0.923)
pooled estimate (2013-2016) = 0.046 (0.035)
p-value joint test of 2013-2016 placebos = 0.4723

(b) Exam score grade 10

sample mean (SD) = 0.036 (0.981)
pooled 2011-2016 estimate = 0.024 (0.034); 2013-2014 = 0.026 (0.038)
p-value joint test of 2011-2016 placebos = 0.6232

(c) Average teacher grades in grade 10

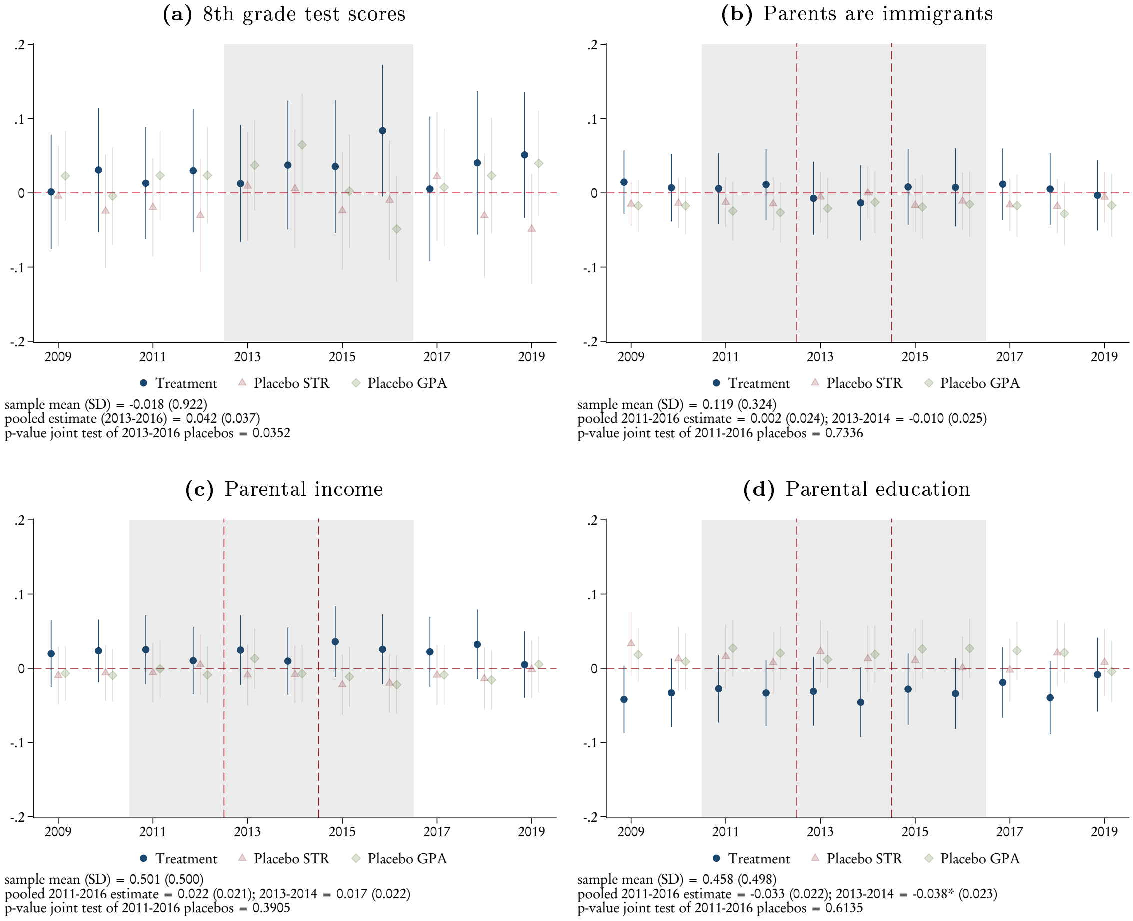
sample mean (SD) = 0.004 (0.992)
pooled 2011-2016 estimate = 0.005 (0.027); 2013-2014 = -0.006 (0.032)
p-value joint test of 2011-2016 placebos = 0.0626

(d) On-time completion of year two of upper seconary

sample mean (SD) = 0.790 (0.407)
pooled 2011-2016 estimate = 0.001 (0.009); 2013-2014 = 0.000 (0.011)
p-value joint test of 2011-2015 placebos = 0.9982

Note: The graphs show estimates and confidence intervals from estimating eq. (1). Outcomes are a) 9th grade test scores, b) exam scores, c) teacher grades and d) completion of year two of upper secondary school. Control variables are year fixed effects. The different series correspond to treatment effects, δ in eq. (1), and placebo effects from the non-treatment margins, γ in eq. (1). The x-axis is the year of the 8th grade test, treated cohorts are shaded. In sub-figures b-d the dashed vertical lines indicate cohorts treated for three years. The figure notes show the sample mean and standard deviation of the outcome, estimated effect and standard errors for a pooled analysis of the treatment years and the p-value of a joint test of all placebo effects (γ) for all treatment years. The regression uses student-level data and clusters standard errors at school level.

Figure A.9. Placebo effects on pre-determined student characteristics



(a) 8th grade test scores

sample mean (SD) = -0.018 (0.922)
pooled estimate (2013-2016) = 0.042 (0.037)
p-value joint test of 2013-2016 placebos = 0.0352

(b) Parents are immigrants

sample mean (SD) = 0.119 (0.324)
pooled 2011-2016 estimate = 0.002 (0.024); 2013-2014 = -0.010 (0.025)
p-value joint test of 2011-2016 placebos = 0.7336

(c) Parental income

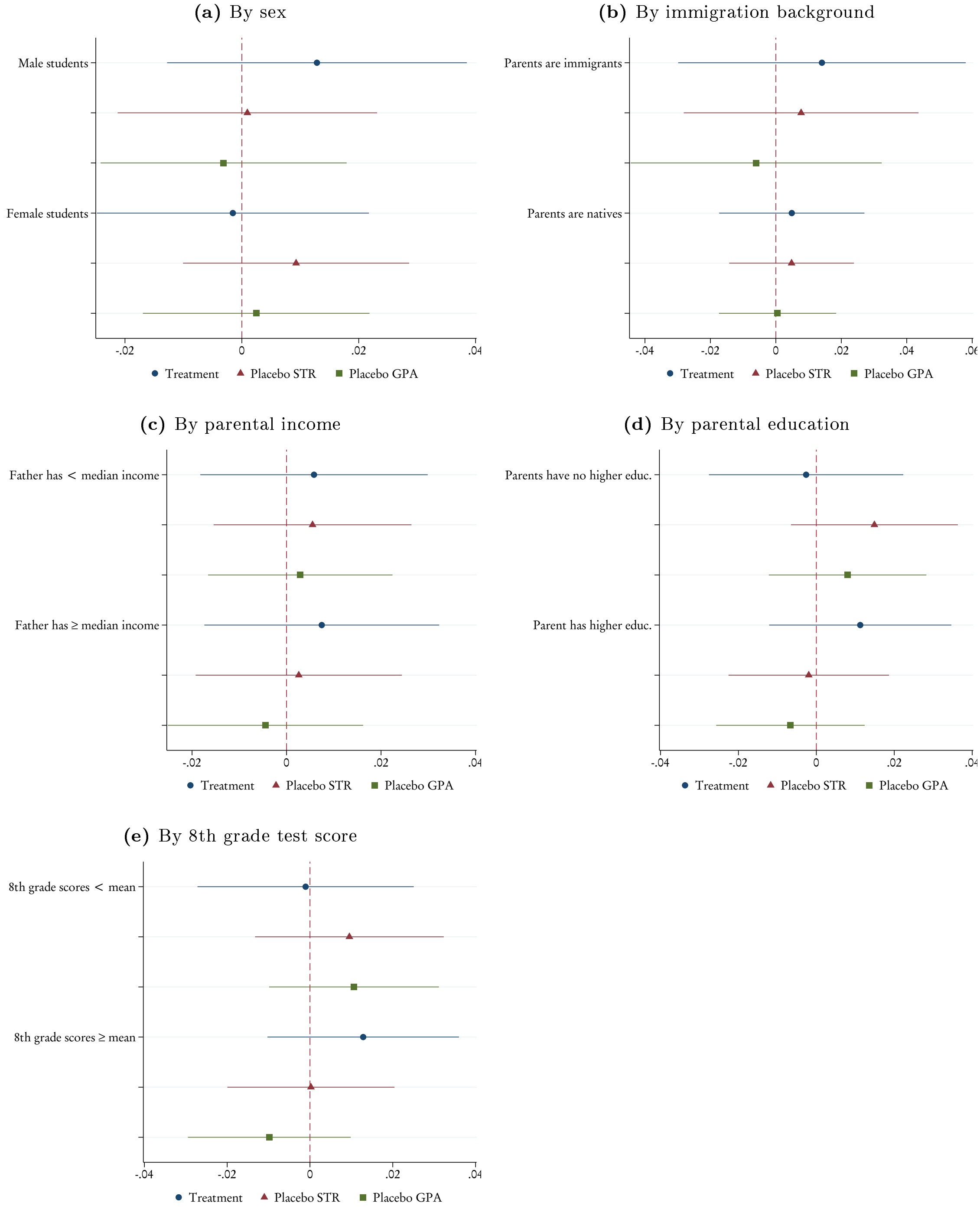
sample mean (SD) = 0.501 (0.500)
pooled 2011-2016 estimate = 0.022 (0.021); 2013-2014 = 0.017 (0.022)
p-value joint test of 2011-2016 placebos = 0.3905

(d) Parental education

sample mean (SD) = 0.458 (0.498)
pooled 2011-2016 estimate = -0.033 (0.022); 2013-2014 = -0.038* (0.023)
p-value joint test of 2011-2016 placebos = 0.6135

Note: The graphs show estimates and confidence intervals from estimating (1). Outcomes are pre-determined student characteristics. The different series correspond to treatment effects, δ in eq. (1), and placebo effects from the non-treatment margins, γ in eq. (1). The treated cohorts are shaded. The dashed vertical lines indicate the cohorts treated for three years. The figure note shows the sample mean and standard deviation of the outcome, estimated effect and standard errors for a pooled analysis of the treatment years and the p-value of a joint test of all placebo effects (γ) for all treatment years. The regression uses student-level data and clusters standard errors at school level.

Figure A.10. Heterogeneous effects on standardized test scores



(a) By sex

(b) By immigration background

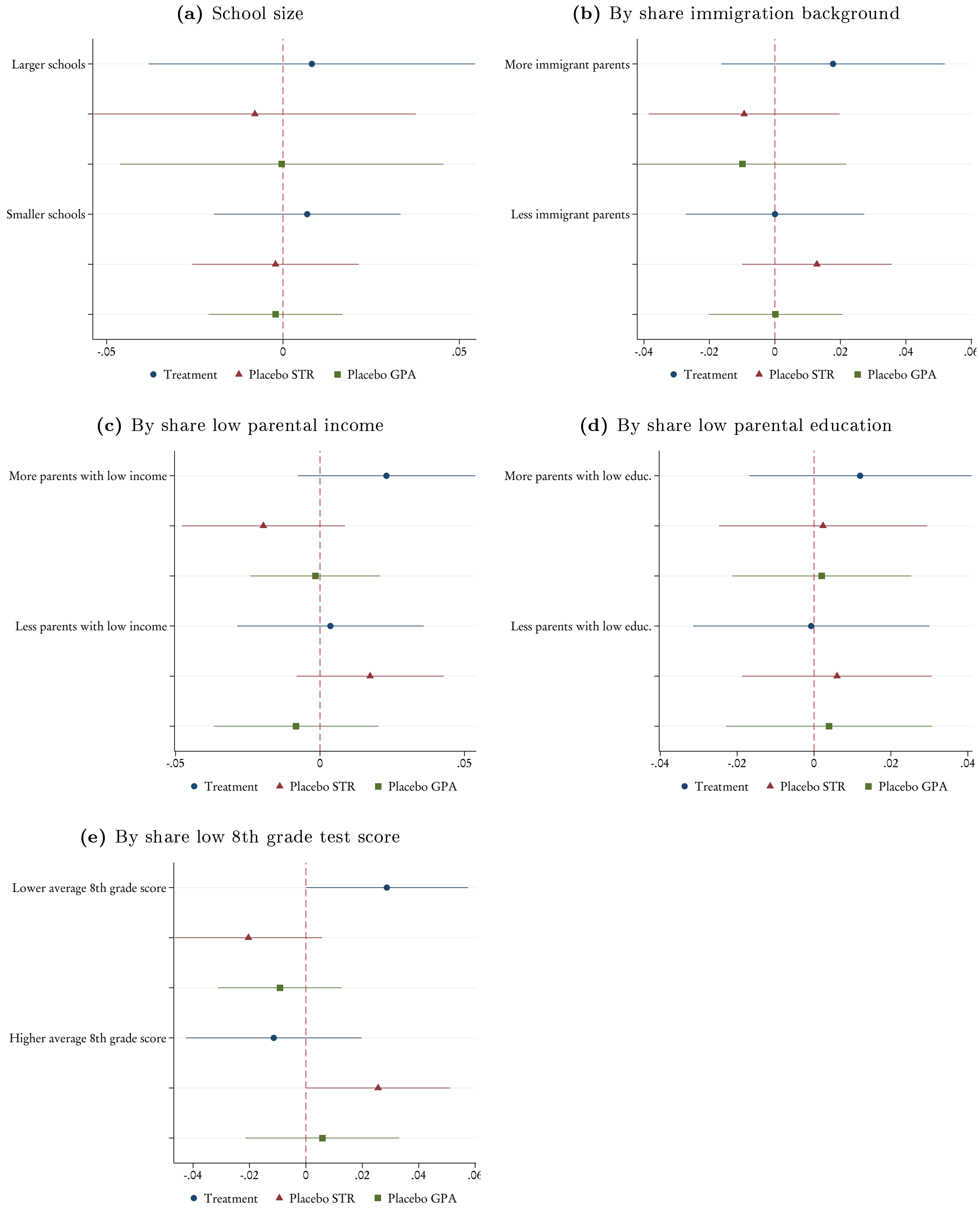
(c) By parental income

(d) By parental education

(e) By 8th grade test score

Note: The graphs show estimates and confidence intervals from estimating (1) fully interacted with a binary variable. Outcome is 9th grade test scores. The different estimates correspond to treatment effects, δ in eq. (1), and placebo effects from the non-treatment margins, γ in eq. (1). The regression uses student-level data and clusters standard errors at school level.

Figure A.11. Heterogeneous effects on standardized test scores, by school characteristics



(a) School size

(b) By share immigration background

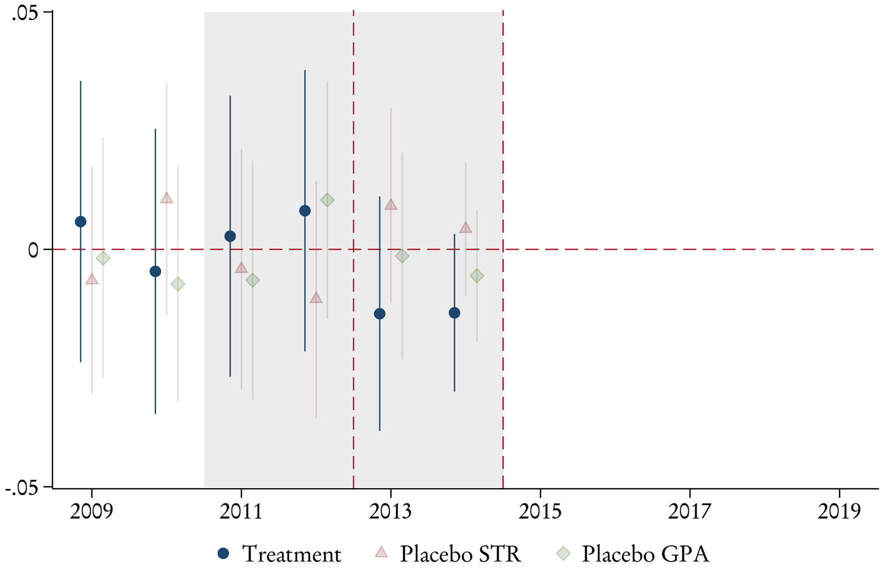
(c) By share low parental income

(d) By share low parental education

(e) By share low 8th grade test score

Note: The graphs show estimates and confidence intervals from estimating (1) fully interacted with a binary variable. Outcome is 9th grade test scores. The different estimates correspond to treatment effects, δ in eq. (1), and placebo effects from the non-treatment margins, γ in eq. (1). The regression uses student-level data and clusters standard errors at school level.

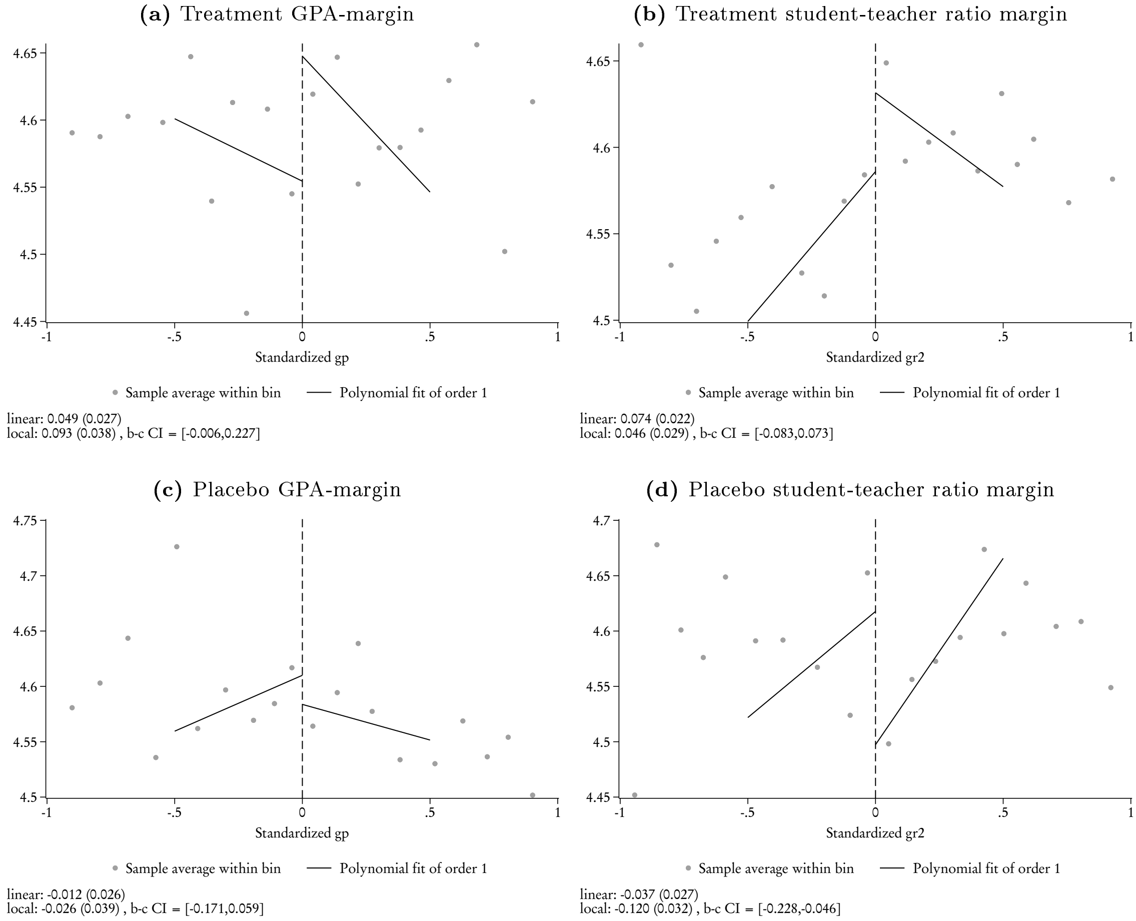
Figure A.12. Effects on enrollment in third year of upper secondary school



sample mean (SD) = 0.737 (0.440)
pooled 2011-2016 estimate = -0.003 (0.009); 2013-2014 = -0.011 (0.008)
p-value joint test of 2011-2014 placebos = 0.4621

Note: The graph shows estimates and confidence intervals from estimating (1). The outcome is completion of the second year of upper secondary school and the control variables are gender, age, year fixed effects, a cubic in the 8th grade test score, and parental education. The different series correspond to treatment effects, δ in eq. (1), and placebo effects from the non-treatment margins, γ in eq. (1). The treated cohorts are shaded. The dashed vertical lines indicate the cohorts treated for three years. The figure note shows the sample mean and standard deviation of the outcome, estimated effect and standard errors for a pooled analysis of the treatment years and the p-value of a joint test of all placebo effects (γ) for all treatment years. The regression uses student-level data and clusters standard errors at school level.

Figure A.13. RD estimates of effects on school environment index



(a) Treatment GPA-margin

linear: 0.049 (0.027)
local: 0.093 (0.038) , b-c CI = [-0.006,0.227]

(b) Treatment student-teacher ratio margin

linear: 0.074 (0.022)
local: 0.046 (0.029) , b-c CI = [-0.083,0.073]

(c) Placebo GPA-margin

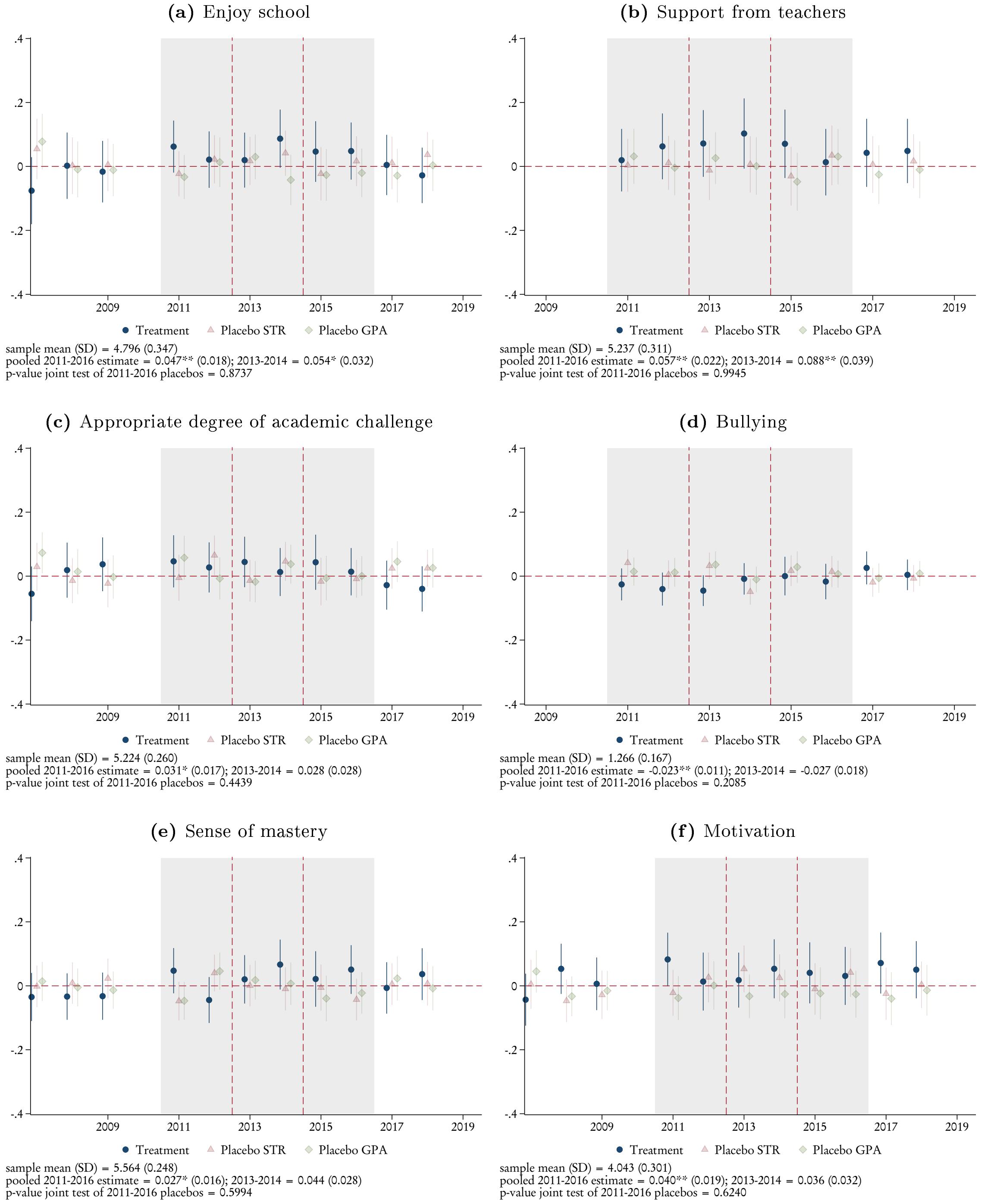
linear: -0.012 (0.026)
local: -0.026 (0.039) , b-c CI = [-0.171,0.059]

(d) Placebo student-teacher ratio margin

linear: -0.037 (0.027)
local: -0.120 (0.032) , b-c CI = [-0.228,-0.046]

Note: The graphs show RD estimates for the school environment index for the years 2013-2016. The graph is based on school-level data and student weights. The lines show the local linear regressions and are estimated using Calonico et al. (2014), with no additional control variables, triangular weights and a fixed bandwidth of 0.5. Bins are quantile-based. Figure notes show coefficients and standard errors from linear and local regressions. Bias-corrected confidence interval (estimated using higher-order polynomial) in brackets.

Figure A.14. School environment sub indices 1



(a) Enjoy school

sample mean (SD) = 4.796 (0.347)
pooled 2011-2016 estimate = 0.047** (0.018); 2013-2014 = 0.054* (0.032)
p-value joint test of 2011-2016 placebos = 0.8737

(b) Support from teachers

sample mean (SD) = 5.237 (0.311)
pooled 2011-2016 estimate = 0.057** (0.022); 2013-2014 = 0.088** (0.039)
p-value joint test of 2011-2016 placebos = 0.9945

(c) Appropriate degree of academic challenge

sample mean (SD) = 5.224 (0.260)
pooled 2011-2016 estimate = 0.031* (0.017); 2013-2014 = 0.028 (0.028)
p-value joint test of 2011-2016 placebos = 0.4439

(d) Bullying

sample mean (SD) = 1.266 (0.167)
pooled 2011-2016 estimate = -0.023** (0.011); 2013-2014 = -0.027 (0.018)
p-value joint test of 2011-2016 placebos = 0.2085

(e) Sense of mastery

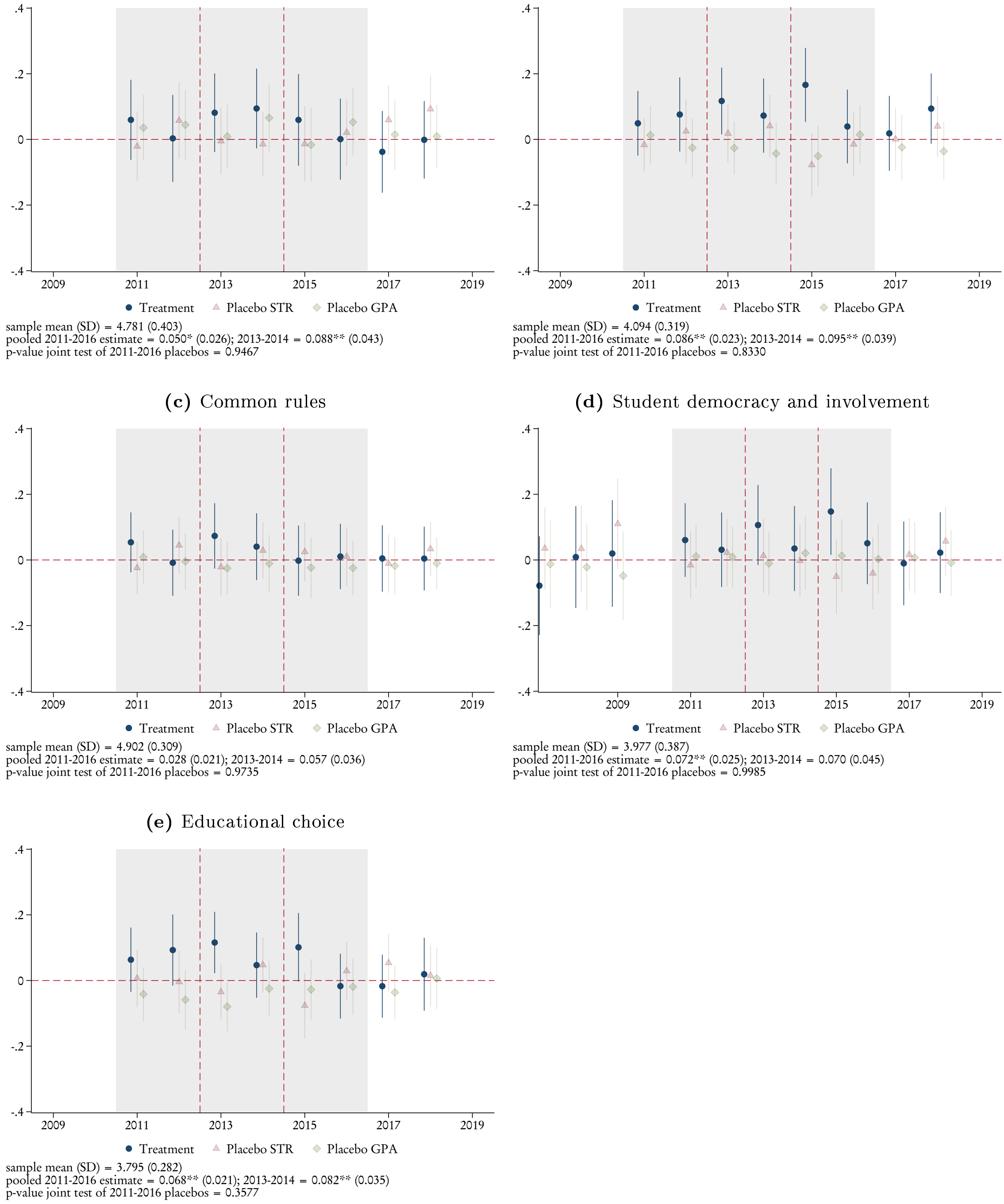
sample mean (SD) = 5.564 (0.248)
pooled 2011-2016 estimate = 0.027* (0.016); 2013-2014 = 0.044 (0.028)
p-value joint test of 2011-2016 placebos = 0.5994

(f) Motivation

sample mean (SD) = 4.043 (0.301)
pooled 2011-2016 estimate = 0.040** (0.019); 2013-2014 = 0.036 (0.032)
p-value joint test of 2011-2016 placebos = 0.6240

Note: The graphs show estimates and confidence intervals from estimating (1). Outcomes indices from a student survey constructed by the Norwegian Directorate of Education and Training. The regression uses school level with student weights. See note to Figure 6 for further details.

Figure A.15. School environment sub indices 2



(a) Learning culture in school

sample mean (SD) = 4.781 (0.403)
pooled 2011-2016 estimate = 0.050* (0.026); 2013-2014 = 0.088** (0.043)
p-value joint test of 2011-2016 placebos = 0.9467

(b) Assessment that supports learning

sample mean (SD) = 4.094 (0.319)
pooled 2011-2016 estimate = 0.086** (0.023); 2013-2014 = 0.095** (0.039)
p-value joint test of 2011-2016 placebos = 0.8330

(c) Common rules

sample mean (SD) = 4.902 (0.309)
pooled 2011-2016 estimate = 0.028 (0.021); 2013-2014 = 0.057 (0.036)
p-value joint test of 2011-2016 placebos = 0.9735

(d) Student democracy and involvement

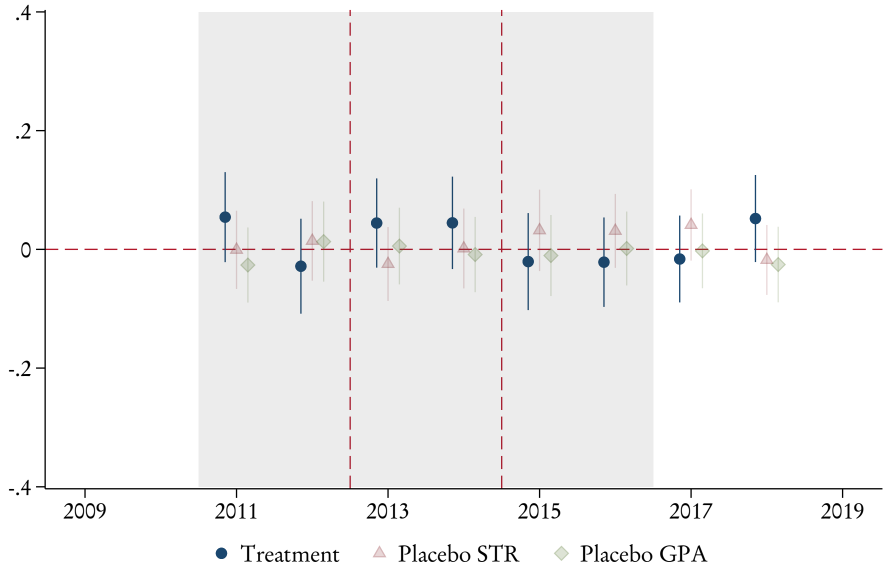
sample mean (SD) = 3.977 (0.387)
pooled 2011-2016 estimate = 0.072** (0.025); 2013-2014 = 0.070 (0.045)
p-value joint test of 2011-2016 placebos = 0.9985

(e) Educational choice

sample mean (SD) = 3.795 (0.282)
pooled 2011-2016 estimate = 0.068** (0.021); 2013-2014 = 0.082** (0.035)
p-value joint test of 2011-2016 placebos = 0.3577

Note: The graphs show estimates and confidence intervals from estimating (1). Outcomes indices from student survey constructed by the Norwegian Directorate of Education and Training. The regression uses school-level with student weights. See note to Figure 6 for further details.

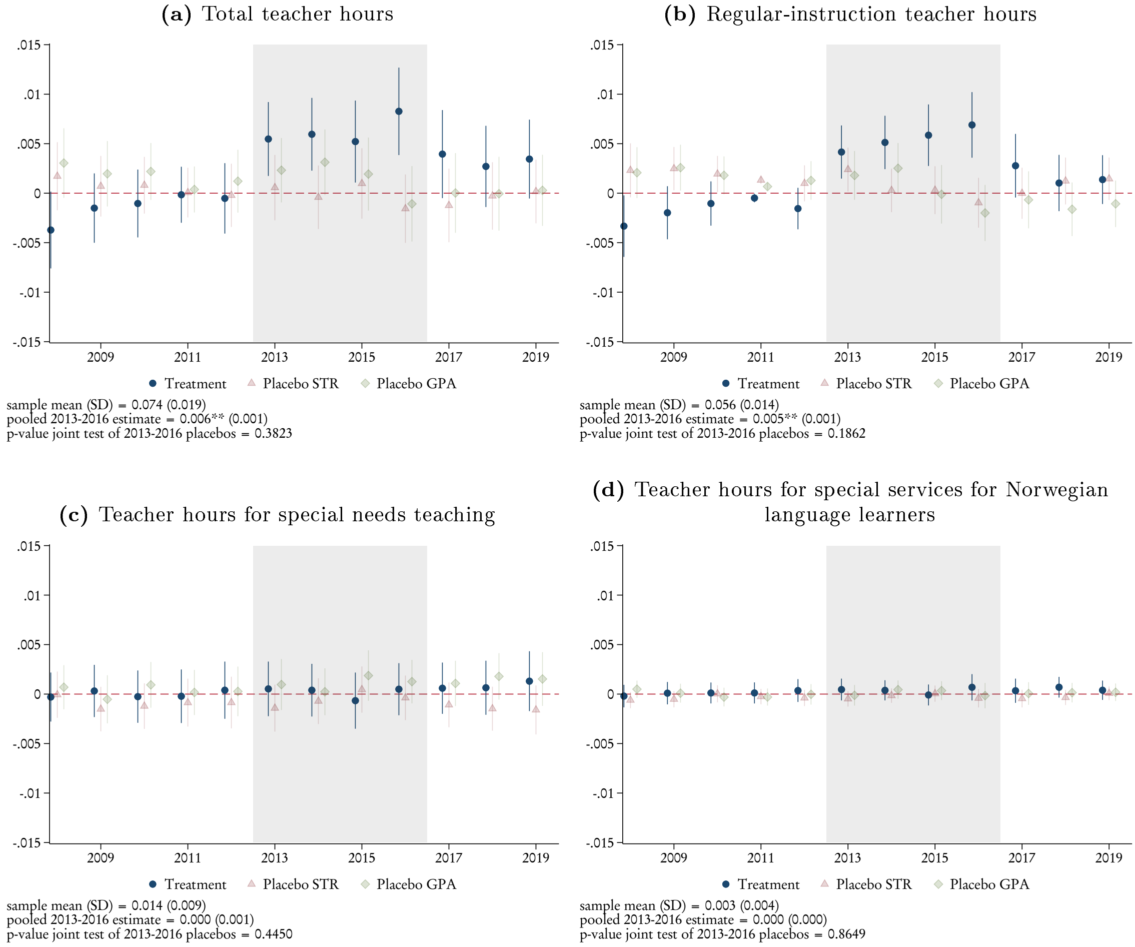
Figure A.16. Support from parents



sample mean (SD) = 4.450 (0.256)
pooled 2011-2016 estimate = 0.012 (0.016); 2013-2014 = 0.044 (0.028)
p-value joint test of 2011-2016 placebos = 0.9766

Note: The graphs show estimates and confidence intervals from estimating (1). Outcomes indices from student survey constructed by the Norwegian Directorate of Education and Training. The regression uses school level with student weights. See note to Figure 6 for further details.

Figure A.17. Effects on total teacher hours and different uses of teacher hours



(a) Total teacher hours

sample mean (SD) = 0.074 (0.019)
pooled 2013-2016 estimate = 0.006** (0.001)
p-value joint test of 2013-2016 placebos = 0.3823

(b) Regular-instruction teacher hours

sample mean (SD) = 0.056 (0.014)
pooled 2013-2016 estimate = 0.005** (0.001)
p-value joint test of 2013-2016 placebos = 0.1862

(c) Teacher hours for special needs teaching

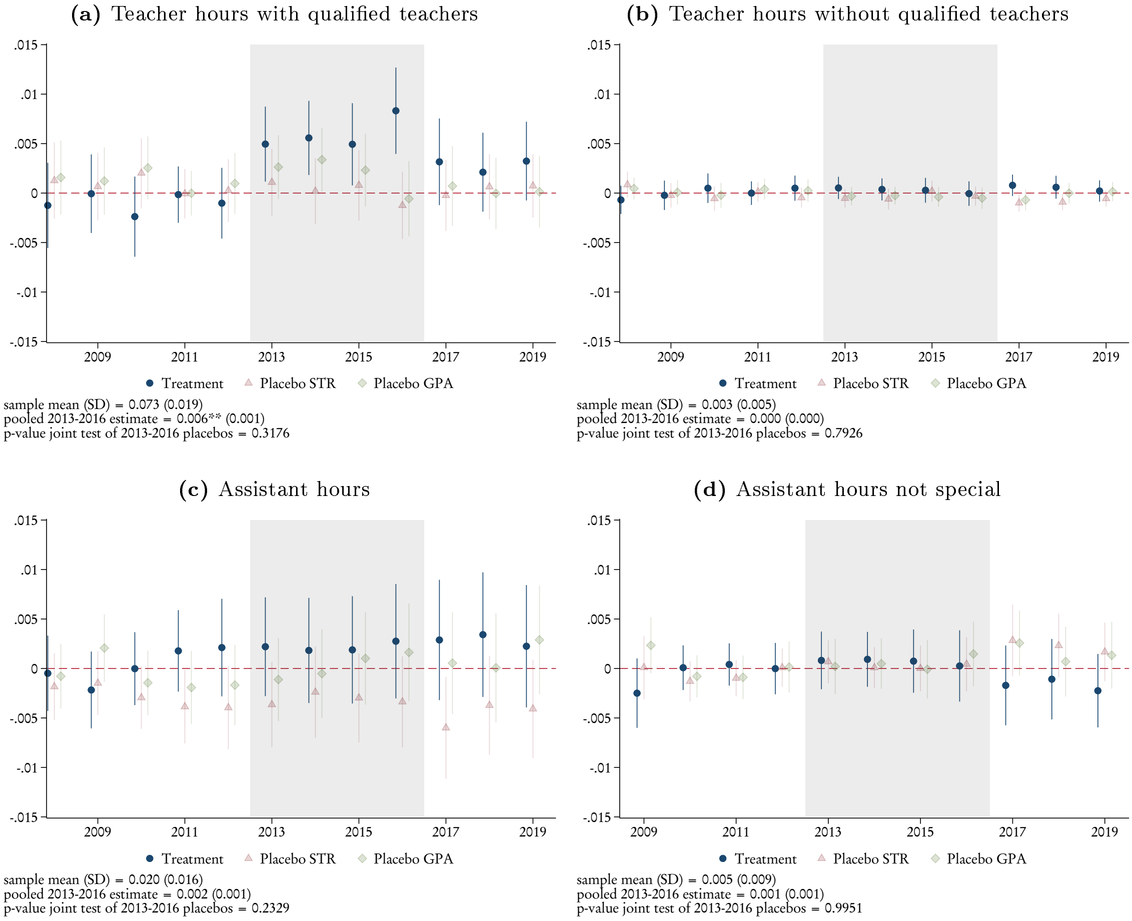
sample mean (SD) = 0.014 (0.009)
pooled 2013-2016 estimate = 0.000 (0.001)
p-value joint test of 2013-2016 placebos = 0.4450

(d) Teacher hours for special services for Norwegian language learners

sample mean (SD) = 0.003 (0.004)
pooled 2013-2016 estimate = 0.000 (0.000)
p-value joint test of 2013-2016 placebos = 0.8649

Note: The graph shows estimates and confidence intervals from estimating (1). Outcomes are total teacher hours and teacher hours decomposed by use, relative to total student hours. The regression uses student weights. See note to Figure 3 for further details.

Figure A.18. Effects on teacher and assistant hours by qualifications and use



(a) Teacher hours with qualified teachers

sample mean (SD) = 0.073 (0.019)
pooled 2013-2016 estimate = 0.006** (0.001)
p-value joint test of 2013-2016 placebos = 0.3176

(b) Teacher hours without qualified teachers

sample mean (SD) = 0.003 (0.005)
pooled 2013-2016 estimate = 0.000 (0.000)
p-value joint test of 2013-2016 placebos = 0.7926

(c) Assistant hours

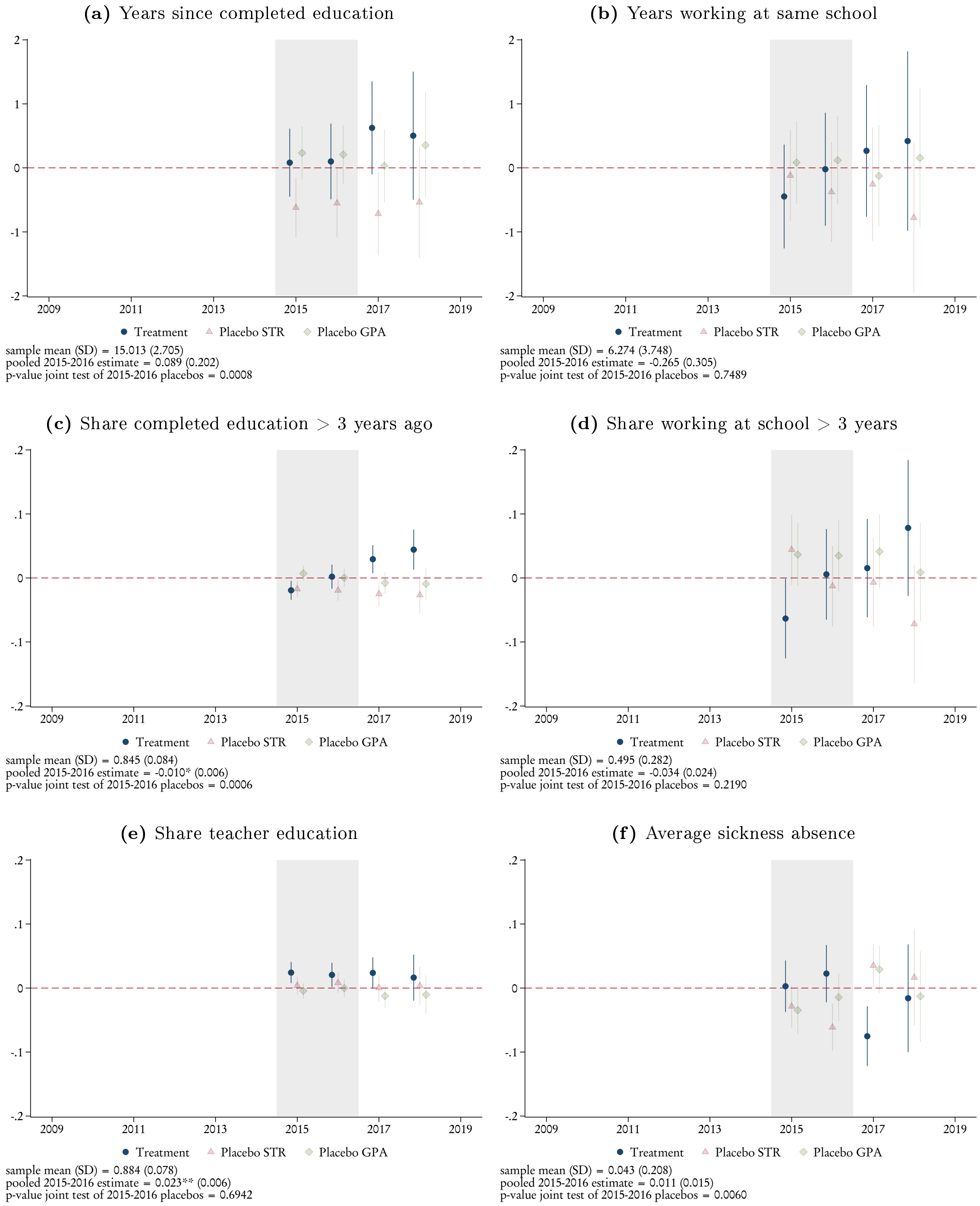
sample mean (SD) = 0.020 (0.016)
pooled 2013-2016 estimate = 0.002 (0.001)
p-value joint test of 2013-2016 placebos = 0.2329

(d) Assistant hours not special

sample mean (SD) = 0.005 (0.009)
pooled 2013-2016 estimate = 0.001 (0.001)
p-value joint test of 2013-2016 placebos = 0.9951

Note: The graph shows estimates and confidence intervals from estimating (1). Outcomes are total teacher hours and assistant hours by qualifications and use, relative to total student hours. The regression uses student weights. See note to Figure 3 for further details.

Figure A.19. Effects on teachers



(a) Years since completed education

sample mean (SD) = 15.013 (2.705)
pooled 2015-2016 estimate = 0.089 (0.202)
p-value joint test of 2015-2016 placebos = 0.0008

(b) Years working at same school

sample mean (SD) = 6.274 (3.748)
pooled 2015-2016 estimate = -0.265 (0.305)
p-value joint test of 2015-2016 placebos = 0.7489

(c) Share completed education > 3 years ago

sample mean (SD) = 0.845 (0.084)
pooled 2015-2016 estimate = -0.010* (0.006)
p-value joint test of 2015-2016 placebos = 0.0006

(d) Share working at school > 3 years

sample mean (SD) = 0.495 (0.282)
pooled 2015-2016 estimate = -0.034 (0.024)
p-value joint test of 2015-2016 placebos = 0.2190

(e) Share teacher education

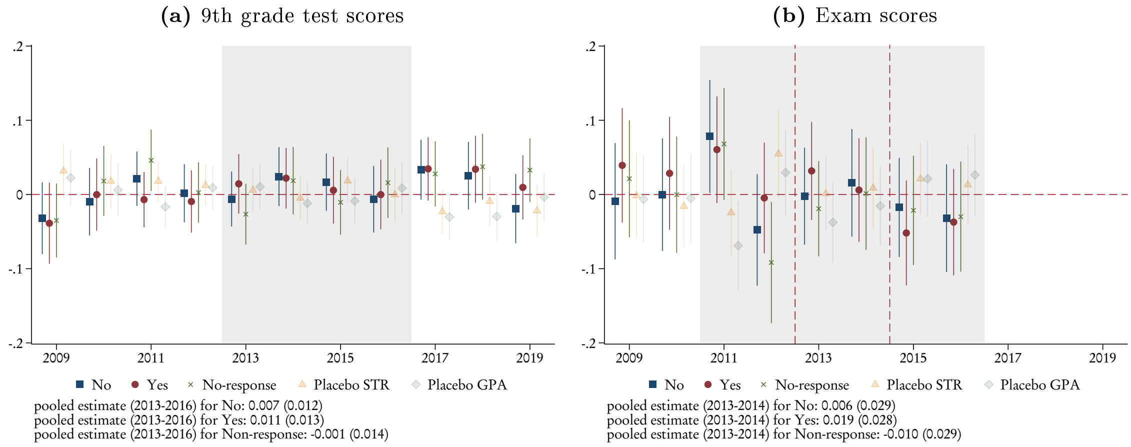
sample mean (SD) = 0.884 (0.078)
pooled 2015-2016 estimate = 0.023** (0.006)
p-value joint test of 2015-2016 placebos = 0.6942

(f) Average sickness absence

sample mean (SD) = 0.043 (0.208)
pooled 2015-2016 estimate = 0.011 (0.015)
p-value joint test of 2015-2016 placebos = 0.0060

Note: The graph shows estimates and confidence intervals from estimating (1). Outcomes are average teacher characteristics derived from matched employer-employee data. The regression uses student weights. See note to Figure 3 for further details.

Figure A.20. Effects on test scores and exam scores by whether treatment involves small-group instruction

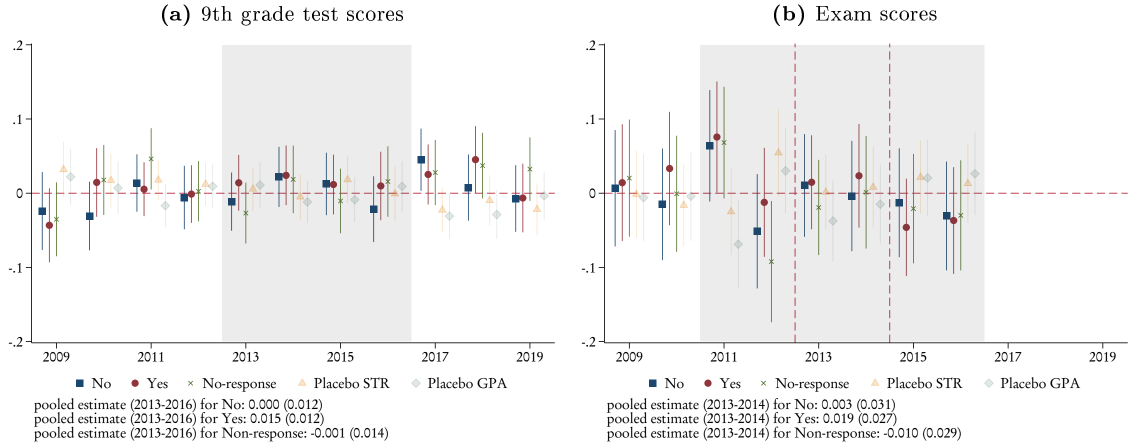


(a) 9th grade test scores

(b) Exam scores

Note: The graphs show estimates and confidence intervals from estimating (1). Outcomes are average 9th grade test scores in figure (a) and exam scores in figure (b). The different series correspond to treatment effects, δ in eq. (1), and placebo effects from the non-treatment margins, γ in eq. (1). The treatment effects are estimated separately for schools according to how the principals answer a survey question about whether the school used the estra teachers for pull-out instruction in small groups (fewer than eight students). The treated cohorts are shaded. The dashed vertical lines indicate the cohorts treated for three years. The regression uses student-level data and clusters standard errors at school level.

Figure A.21. Effects on test scores and exam scores by whether treatment involves ability grouping

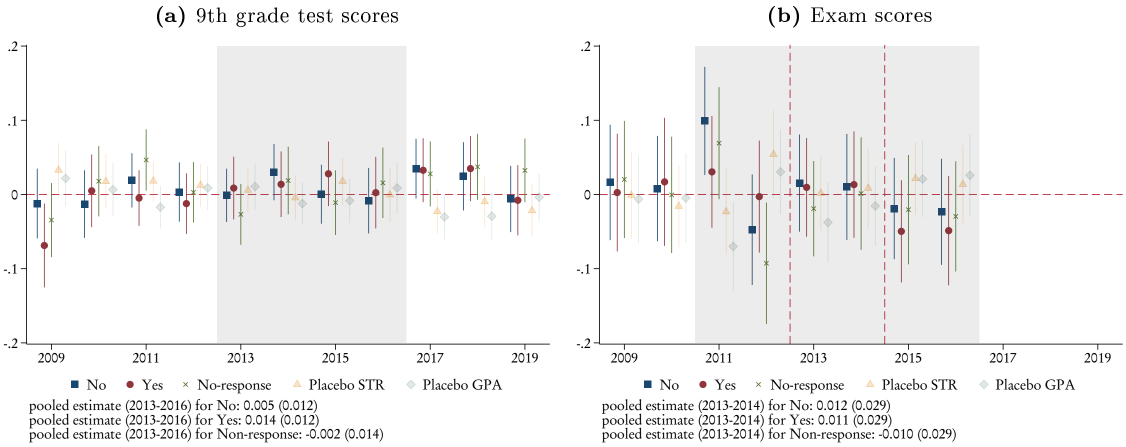


(a) 9th grade test scores

(b) Exam scores

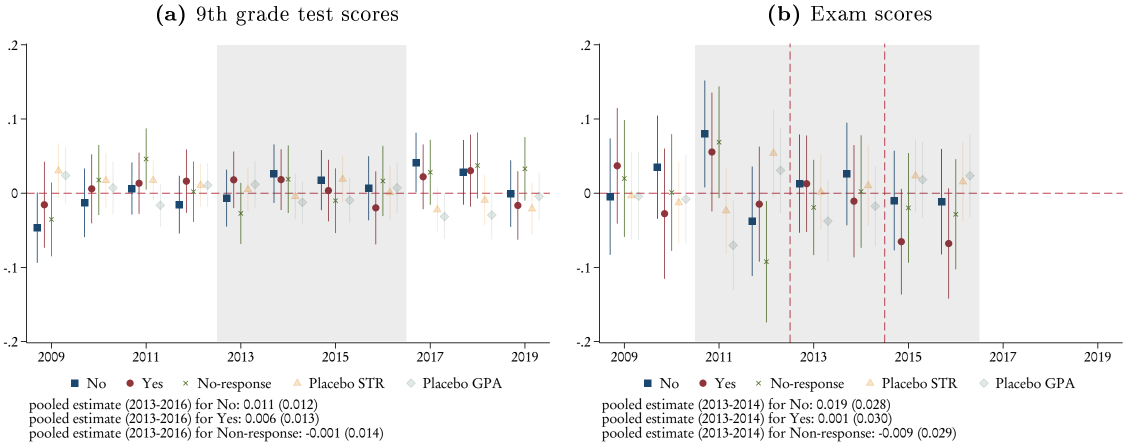
Note: The graphs show estimates and confidence intervals from estimating (1). Outcomes are average 9th grade test scores in figure (a) and exam scores in figure (b). The different series correspond to treatment effects, δ in eq. (1), and placebo effects from the non-treatment margins, γ in eq. (1). The treatment effects are estimated separately for schools according to how the principals answer a survey question about whether the school used the estra teachers for ability grouping (for shorter periods of time, ie., streaming). The treated cohorts are shaded. The dashed vertical lines indicate the cohorts treated for three years. The regression uses student-level data and clusters standard errors at school level.

Figure A.22. Effects on test scores and exam scores by whether treatment involves prioritizing the lowest-performing students



Note: The graphs show estimates and confidence intervals from estimating (1). Outcomes are average 9th grade test scores in figure (a) and exam scores in figure (b). The different series correspond to treatment effects, δ in eq. (1), and placebo effects from the non-treatment margins, γ in eq. (1). The treatment effects are estimated separately for schools according to how the principals answer a survey question about whether the school used the estra teachers primarily to focus on the lowest-performing students. The treated cohorts are shaded. The dashed vertical lines indicate the cohorts treated for three years. The regression uses student-level data and clusters standard errors at school level.

Figure A.23. Effects on test scores and exam scores by whether treatment involves prioritizing mathematics teaching



Note: The graphs show estimates and confidence intervals from estimating (1). Outcomes are average 9th grade test scores in figure (a) and exam scores in figure (b). The different series correspond to treatment effects, δ in eq. (1), and placebo effects from the non-treatment margins, γ in eq. (1). The treatment effects are estimated separately for schools according to how the principals answer a survey question about whether the school used the estra teachers in most mathematics lessons. The treated cohorts are shaded. The dashed vertical lines indicate the cohorts treated for three years. The regression uses student-level data and clusters standard errors at school level.