Towards Evidence-Based School Accountability in Sweden

Axel Norgren

Institute for Analytical Sociology, Linköping University, Sweden; Center for Educational Leadership and Excellence, Stockholm School of Economics, Sweden
axel.norgren@liu.se
+46703625711

Abstract: School accountability is a common feature of many educational systems, with public school inspections common in European countries. Yet, there is scarce empirical evidence on the inspectorate model of school accountability. This article evaluates the effects of Swedish school inspections on pupils' achievement using a double-robust difference-in-difference model. The results show overall negligible effects of inspection but with indications of effects heterogeneity over time. Such short- long-term outcomes should be considered in future research and educational policy when it comes to school accountability.

1. INTRODUCTION

Many nations have seen the quality of their educational systems decline in international assessments, and policy makers are grasping for measures to improve their national educational performance. A popular call to action is school accountability. School accountability as a measure can be designed as a top-down government system with national programs, agencies and interventions that are meant to evaluate and monitor schools. This assessment can be done against standards of educational quality and other values to ensure accountability towards the stakeholders. As the accountability movement is expanding globally it now comes in very different shapes and forms.

The school accountability movement has in a European setting mostly taken the form of school "inspections" where the British Office for Standards in Education (Ofsted) has been an important reference point in relation to the foundation of other school inspectorates (SOU, 2007:101). School inspections as an accountability mechanism comes with empirical variation across education systems and time, but usually contains some sort of visit by public inspectors, systematic evaluation and feedback to schools (Ehren et al., 2015; Hofer et al., 2020). The school inspection model does not necessarily only focus on schools' educational performance, it is common that legal or organizational aspects of the schools are assessed as well. Most school inspection systems, however, share the goal of improved educational standards with the broader school accountability movement. Empirical support for the effectiveness of school accountability and school inspections on educational output remains mixed. The international accountability literature findings of small to modest positive effects in mostly a US context (Figlio & Loeb, 2011; Figlio & Ladd, 2014; Hanushek & Raymond, 2005) of large-scale accountability programs. The literature on the inspection model of school accountability, however shows a variability in the results with a majority finding no effect of school inspections on pupils educational achievement (De Wolf & Janssens, 2007; Hofer et al., 2020; Hussain, 2015; Luginbuhl et al., 2009; Rosenthal, 2004).

One reason underlying the mixed results in studies of school inspection effects is the challenge of causal inference in school inspection settings. The literature on the inspection model of school accountability has struggled with endogeneity problems given the observational nature of the data (and assignment mechanisms) through either fixed effects models (Luginbuhl et al., 2009; Rosenthal, 2004), difference-in-difference estimation (Hussain, 2015) or by randomization in assignment of inspection (Luginbuhl et al., 2009). While fixed effects models are good at purging time-consistent factors, they can not capture

time-varying unobservable heterogeneity. Difference-in-difference models have the additional trait that common time-varying factors are captured by the research design. As the difference-in-difference methodology is constantly developing, new designs are available that provide flexible frameworks to study observational phenomena with more credible design (Callaway & Sant'Anna, 2021; Roth et al., 2023; Sant'Anna & Zhao, 2020).

The state-of-the-art difference-in-difference methodology is robust (doubly so) to model misspecification (Sant'Anna & Zhao, 2020) and can be combined with event-study designs to examine effects across multiple post-treatment time periods. As studies on school inspections cover post-inspection effects for up to 4 years (Hussain, 2015; Luginbuhl et al., 2009; Rosenthal, 2004) this study will be able to study the effects of a school inspection over a 6-year period[1] with an application to Swedish data. This is suitable given the possibly differential effects found in the short- vs the medium term in previous studies (Rosenthal, 2004). The currently Swedish dataset also allows us to exploit an inspection gap at the time of the Covid-19 pandemic to examine the effects of an inspection visit over a longer time period.

The empirical body of evidence does not justify the resources being spent on school accountability at this date. The literature on the European model of school inspection is sparse and mostly looks at short-term effects. Given the recent methodological advance, it is warranted to demonstrate how modern difference-in-difference methodology can be applied to derive dynamic effects of school inspections over a longer time-horizon. As the positive findings in the U.S literature on school accountability contrast the mixed effects of the European literature, it is important that separate empirical foundation for the European system of school inspection is expanded.

This leads to the primary research question:

Q1: Do the Swedish School Inspectorates inspection visits increase educational outcomes?

This study provides several contributions to the school accountability literature. The first contribution is that it is the first study utilizing double robust difference-in-difference model to study the effects of a school inspection visit for students' educational achievements. It will also be the first to examine the effects of a school inspection visit on educational outcomes over a longer time horizon (6-year period). It will also be one of few studies studying the

[1] 5 years post treatment years, but 6 years including the year of treatment

effects of the European inspectorate model of accountability and exclusively on the Swedish inspection model on educational outcomes.

As the empirical body of evidence of the inspectorate model is scarce, the novel approach (in this context) offers a way to identify valid treatment effects relying on two strategies of weighting by and adjusting for covariates, hence it is "doubly robust". This allows for credibly estimating treatment effects in a difference-in-difference setting as it is sufficient that one of two strategies is correctly specified.

The results of this study show that the net effect of school inspection varies in the short and long term. First, we find a positive effect on grades for the year 2021. However, this could be due to increased grading leniency for that year due to the abolishment of the grade benchmark (the national exams) during the Covid-19 pandemic. The results also indicate short-term negative effects on test scores the year of inspection, but positive longer-term effect of test scores in mathematics 5 years post-inspection. In sum, the results are somewhat contradictory but also too unstable to make any conclusion of the net effect of a school inspection. The heterogeneity in effects over time suggests the importance to distinguish between long- and short-term effects when it comes to studying effects of school accountability interventions.

The remainder of the paper is organized as follows. Section 2 deals with the contextual background of school accountability and school inspections while also reviewing the empirical evidence. The section also explains the Swedish School Inspections operations. Section 3 describes the theoretical expectations based on the proposed mechanisms. Section 4 describes the data, methodology and identification strategy. Section 5 displays the results with corresponding robustness tests and heterogeneity analysis. Section 6 discusses the finding and section 7 concludes.

2. BACKGROUND

The school accountability movement

The modern idea of school accountability originates from the mid to late 1900s as partly a political product in the U.S. From the 60s with the "war on poverty" and the debate regarding the Coleman report, to the economic stagnation in the 70-80s, smaller educational reforms toward school accountability were taken through increase federal involvement in the otherwise local school system (Hayes, 2008). In the 1980s when what today is known as "new public management" was introduced in the public sector (Figlio & Ladd, 2014). The primary

economic motivation was to measure and evaluate the performance of the public sector similarly to how a private business is operated to increase the effectiveness of the focal organization. Another economic rationale behind school accountability is the principal-agent problem: schools could behave contrary to the interests of the stakeholders (Figlio & Ladd, 2014; Jacob, 2017). For example, parents and politicians might wish for a school to implement and enforce a certain rule such as phone-free classrooms which the school potentially could deviate from. School accountability thus builds on the notion that schools need to be monitored and incentivized to act in the best interest of the stakeholders. Even as school accountability has existed for a longer time than the 1980: s a paradigm formed around this time and culminated in the post-cold war environment when the "No Child Left Behind" (NCLB) act was introduced by the Bush administration and ratified in 2002.

A hope was that when the NCLB was introduced, it would decrease educational inequalities and increase the general level of education in the U.S. An early study of the different accountability measures taken in the U.S was done by Hanushek and Raymond (2005). Their findings did indicate that accountability systems were generally successful in improving student achievement but only if there were tangible consequences attached to the specific measures, such as monetary reward or fines. Also, they found adverse effects for inequality such that school accountability widened the white-black student achievement gap. A review of the literature by Figlio and Loeb (2011) note that most school accountability studies tend to show small but positive effects (around 0 to 0.2 standard deviation) for students' educational performance. They also find that studies show mixed effects on ethical subgroups while low-performing schools seem to benefit from school accountability. A later review by Figlio and Ladd (2014) further supports these results but also highlights that the positive effects of school accountability seem to be larger in mathematics, plausibly as a consequence of the high classroom intensity of the mathematics subject. This review also notes that accountability systems easily result in schools engaging in "gaming" of the educational systems, with different adverse consequences could happen due to maximization of educational achievement such as teaching to the test, manipulation of student groups and outright cheating (Jacob, 2017). Recent literature also shows long-term positive effects of school accountability measures such as the NCLB programs on high-school graduation rates (McElroy, 2023).

The School Inspectorate model

While the U.S. accountability systems tend to focus more narrowly on educational performance indicators such as student test scores, European model of school accountability usually take a broader approach. Beyond looking at educational performance, the school inspection model usually also focuses on school environment, staff and educational practices (Hofer et al., 2020). Even as the inspectorate model of school accountability differs in its practice across countries, the core fundament is regular visits to schools and evaluating the schools against various legal and process quality standards (Ehren et al., 2015; Hofer et al., 2020).

A focal point of the modern school inspection is the English Office of Standards in Education (Ofsted) which has existed since 1992. Ofsted performs cyclical visits to schools, holding schools accountable on legal, educational and performative standards while also publicly giving the schools an assessment if they "failed or not". An early study into the effectiveness of Ofsted's inspection visits on exam performance by Rosenthal (2004) used fixed- and random effects panel models to analyze 2300 English secondary schools. Rosenthal (2004) found no effects on exam performance the year after inspection but an adverse effect in the year when inspections took place, an effect possibly stemming from school staff being distracted by meeting Ofsted's standards and benchmark rather than focusing on teaching. The Dutch school inspectorate is procedurally similar to Ofsted. A study on the Dutch School Inspectorate by Luginbuhl et al. (2009) finds contradictory results to Rosenthal (2004) with fixed-effects models, with student test scores in inspected schools increasing by 2-3% of a standard deviation 2 years after the inspection. The results also appear stronger for mathematics test scores, similar to findings from U.S school accountability studies. A more recent study on the effects Ofsted's visits and ratings by Hussain (2015) employing a difference-in-difference design found positive effects on primary school pupils test scores (mathematics and English) in schools that "fail" an inspection, especially for the students that are previously low-scoring on tests. However, simply being inspected and passing the inspection appears to have no effect (Hussain, 2015). The results are also robust against "gaming behavior" (Hussain, 2015).

Two recent review articles cover most studies on the effects of school inspections. De Wolf and Janssens (2007) review the literature on the effects and side effects of European control systems in education. Overall, they find it is difficult to draw any scientific conclusions about the effectiveness of school inspections, with results differing between studies depending on

analytical method employed, characteristics of the inspection system, and other contextual factors. They also note a lack of studies on the potential side effects of gaming the system. A more recent review by Hofer et al. (2020) that systematically reviews 30 years of research on school inspections finds that a majority of estimated effects (58 %) are not statistically significant from zero. The review only includes studies based on statistical inference, which can vary in the credibility of their research design. Also, this review includes studies that study other outcomes than educational achievement. It also confirms that most positive effects tend to be found for measures of student mathematics achievements, while most negative effects tend to be found for measures of "instructional processes" as the outcome variable.

The Swedish School Inspectorate

The Swedish School Inspectorate (SSI) was instituted in 2008 to enhance the educational quality in the Swedish school system while holding the school accountable to legal standards (SOU, 2007:101). The main purpose of the SSI inspection activity is to "contribute to enhanced achievement, quality and equality" in Swedish schools. SSE express that they "put demands on school organizers to follow the law, develop their organizations, improve and guarantee high quality" (The Swedish School Inspectorate, 2020b). The SSI was partly developed with Ofsted as an archetype (SOU, 2007:101) but with the important distinction that schools are rarely assessed against their educational performance nor do they receive any rating (Ehren et al., 2015). Otherwise, SSI and Ofsted are quite similar with a cyclical visit and assessing educational practices as well as the legality of operations. In fact, the SSI has been heavily criticized for one-dimensionally focusing on the legal aspects of education (Alvesson & Strannegård, 2021). Over time the SSI has gotten more mandate in terms of the ability to issue sanctions to schools (in 2011) and to revoke school's licenses (in 2018) (Ramböll, 2023). The SSI also has prior given notice to schools before they visit (roughly 3 weeks' notice) but has gotten a new directive to show up unannounced (The Swedish School Inspectorate, 2023).

Even as the SSI are tasked with self-evaluation (SOU, 2007:101), they have not previously been evaluated against the benchmark of educational achievement in the schools they visit. This is quite counterintuitive, given the original purpose behind the organization's existence and undoubtedly still the core of their tasks. However, a number of descriptive reports suggests that the SSI have a positive impact on the schools educational processes (Gustafsson, 2014; Strategirådet, 2016; The Swedish School Inspectorate, 2017, 2018, 2019), most relying

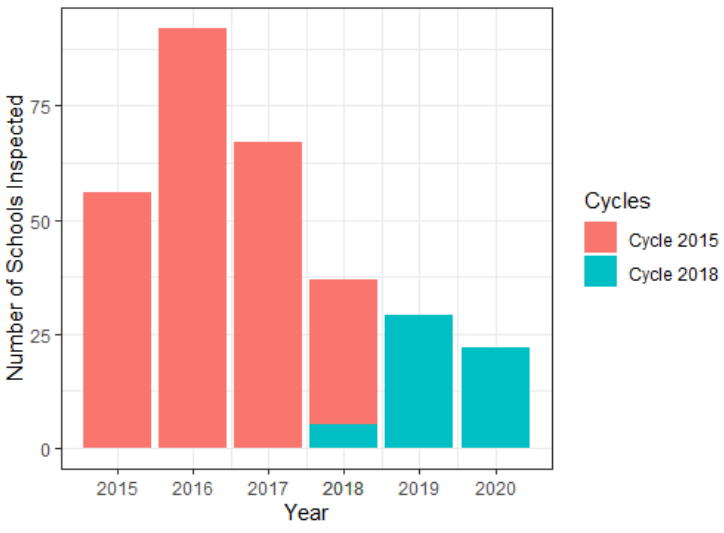
on survey-data and interviews. As discussed in Hofer et al. (2020) self-reported instruments when it comes to school inspections can be biased through social desirability and personal stakes, when answering surveys under accountability pressure.

The Inspection logic

At the onset of their inspection activity, SSI started by inspecting all Swedish elementary, lower- and upper secondary schools between 2009-2014 (corresponding to roughly 5600 school decisions). After this, SSE switched its inspection process to a 'multi-year inspection cycle' that began in 2015 and concluded in the summer of 2018, where focus was on the schools "with the highest development needs" (The Swedish School Inspectorate, 2015). After the summer of 2018, the SSI changed their inspection model again with a narrower focus (targeted only the 5 percent most problematic schools) while allowing for more flexibility (The Swedish School Inspectorate, 2018).

Regarding the cyclical nature of the inspections, Figure 1 below displays the number of public schools that received an inspection over the years 2015-2020[2]. There is a clear trend toward fewer inspection visits partly due to the SSI changing its inspection structure with the cycle that started in the summer 2018 but also due to the Covid-19 pandemic.

Figure 1: Number of schools inspected over different cycles



[2] No public schools were inspected 2021-2022

As mentioned, SSI's inspections are done in cycles, and their methodology has changed over time. Here, I will give a brief description of the logic and methodology of the 2015 cycle that will be the focus of the evaluation in the paper. The logic behind the 2015 cycle was to inspect schools with the "highest risks that the students do not get the education they are entitled to" (The Swedish School Inspectorate, 2015). This motive for inspections is formulated with a focus on students rights which is tightly linked to the legal standard the schools are to be assessed against (The Swedish School Inspectorate, 2024b). Schools are selected for inspection through a risk-index which is a composed of: survey responses regarding school work environment and schools leadership, student educational outcomes, complaints against the school as well as relevant information from previous inspections (The Swedish School Inspectorate, 2015, 2020a).

The inspections are conducted as the following: A few weeks prior to inspection, the school is notified that an inspection will be taking place and asked to submit certain documentation and answer some questions (The Swedish School Inspectorate, 2024a). The school is then visited by inspectors that do classroom observations and interview students, teachers and the principal. After this the SSI submit a public assessment of the school that includes all deficiencies of the school. The school is then reprimanded to fix their deficiencies within a given timeframe and is otherwise threatened by monetary fines by SSI (The Swedish School Inspectorate, 2024a). According to the SSI, the inspection visits can vary in how many and long classroom observations that are done and who is interviewed at the school (The Swedish School Inspectorate, 2017).

Regarding the assessment of the schools, it is in its core legal. The SSI:s formalized assessment list only contain items with a reference to a Swedish School law in it (The Swedish School Inspectorate, 2024b). The inspections look at different themes all anchored in the law, in the 2015 cycle these were *Teaching and Education*, *Adaptability and Support for the Student, Safety and Work Environment, Grading, Prerequisites for Learning and Safety* and finally *Management of the Organization* (The Swedish School Inspectorate, 2017).

3. THEORETICAL MECHANISMS

There are several potential mechanisms at play in how school inspections could affect educational outcome, these might also work at different time horizons.

First off, the schools might become incentivized to organizationally improve when there is a threat of sanctions (Figlio & Loeb, 2011; Figlio & Ladd, 2014). This mechanism does not necessarily need to be tied to inspection itself but also the mere existence of the threat of inspection (pre-emptive effect). Hence this study would not capture this mechanism fully but only partly as it can be believed that the threat of sanction would increase after an inspection visit. This mechanism could of course be related to the principal-agent economic rationale where the inspection as a control mechanism creates a focus toward certain goals that the principal considers more important than the agent, such as learning outcomes. As organizational change is unlikely to happen quickly for schools (Meyers & Smylie, 2017; Murphy & Meyers, 2007) a medium to long time-horizon is expected to be at play here if any positive effect of inspection would work through this mechanism. This effect is likely to be moderated by both student SES and subject since vulnerable schools (Figlio & Loeb, 2011; Hussain, 2015) seem to moderate the relationship as well as the mathematics subject (Figlio & Ladd, 2014; Hofer et al., 2020; Luginbuhl et al., 2009). Potential explanations for this could be that vulnerable schools with low SES are also the ones where an inspection has the largest marginal effect if the organization is more dysfunctional than schools with higher SES. The mathematics subject is theorized to be more "classroom intensive" thus more likely to be impacted by an organizational change at the school.

Secondly, other mechanisms can be the window dressing and "gaming the system". (Figlio & Ladd, 2014; Jacob, 2017) of the school in conjunction with the inspection visit. This can also be tied merely to the threat of inspection but would be expected to intensify when a school is inspected. If a system is governed by an outcome (i.e. a standard) there will be incentives to achieve this outcome in the most efficient manner, albeit through cutting corners or not. Thus, if an outcome can be achieved by having the right documents ready once in a blue moon, then why enforce a costly policy year in, year out? It is partly the new public management rationale that creates these adverse effects. Given the multiplexity of schooling as a service, it is difficult to set up standards for all its intended purposes and doing so in strategy proof manner. As found in Rosenthal (2004) inspections could have short-term negative effects as the school is distracted from its core operations by the inspection to fulfill a standard in the short term. Hence, any negative potential side-effect from inspection is likely to be found in the short to medium term.

Finally, as the context are important for the study at place the Swedish school inspection has been criticized for one-dimensionally focusing on legal standards (Alvesson & Strannegård,

2021) but also shown to improve educational processes according to surveyed principals (Gustafsson, 2014). The literature on the Swedish context is conflicting and does not help in establishing any stronger mechanisms beyond the already mentioned.

4. METHODOLOGY

Data

The data utilized for this study covers the years 2013-2022 and stems study stems from a variety of sources but primarily from the Swedish National Agency for Education and the Swedish School Inspectorate. The main treatment variable is if the school has received an inspection visit or not which is collected from the Swedish School Inspectorate. The outcome variables are *Grades,* the National *Mathematic test* and the National *English test*. These are all collected from the Swedish National Agency for Education. The *Grades* variable is a continuous scale representing the mean accumulated grade points for all subjects in the school at the end of year 9. The national tests are the mean scores at the annual national test for grade 9 of the school taken during the spring. The data for the national exams do not cover 2020-2021 as it was cancelled due to Covid-19. It neither does cover 2018 for the national *Mathematics test* as it the results were revoked due to a major cheating scandal (SVT, 2018). Other variables used as covariates in the outcome model include the variable *parents education* which is measured as the average educational level of the students' parents (it is scored with 1, 2 and 3 where 1 corresponds to completion of primary school, 2 corresponds to completion of lower secondary school and 3 corresponds to completion of post-secondary education). The variable *immigrants* indicate the percentage of students at the school who have been registered in Sweden in the last four years. The variable share of *boys* indicated the percentage share of boys in the school. These variables are taken from the Swedish National Agency for Education.

There is also cross-sectional data utilized for the analysis. The principal's *leadership* is taken from the teacher’s survey where they answer if they think the principal, is a good pedagogical leader. *Safety* is taken from the student’s survey where they answer if they think the school has a safe and calm study environment. Here both *leadership* and *safety* are taken from the School Inspectorate's school surveys. Here, the index of the question category is used with

values between 0–10. For the cross-sectional data, a variable on the number of *complaints* was also utilized which represents how many complaints the SSI received about the focal school. Finally, we also used variables on *students per teacher* and share of *certified teachers* from the Swedish National Agency for Education.

Empirical strategy

To study the effects of a school being inspected on educational outcomes we opt to implement a doubly robust difference-in-difference (Sant'Anna & Zhao, 2020). The difference-in-difference estimator is commonly used as it is effective at mitigating time-constant factors that might distort the analysis but also common time-varying factors. However, its crucial assumption is that of parallel trends in the variable of interest between the treatment and control groups. Given that the treatment in question is not random it is possible that the parallel trends assumption (PTA) in this study does not hold. However, it's here the double robust procedure is useful. The key idea in the model is that it combines the technique of controlling the development of the outcome variables with covariates (regression) while also using covariates to construct weights (IPW i.e Inverse Probability Weights) that weight the model. These approaches are both used to meet the PTA conditional on the covariates. Hence the model shares the strength of both approaches and as long as one of these approaches is correctly specified it allows for identifying the average treatment effect on the treated (ATT).

The model equation utilized looks like the following:

$$y_{it} = \sum_{t=2014}^{2022} \beta_t Year_t * Inspection_i + \beta_2 X_{it} + \varepsilon_{it}$$

The equation above represents our main specification with y_{it} indicating our outcome variable (*Grades* or national tests in *Mathematics* or *English* expressed as standard deviations) for school i in time t . These educational outcomes are chosen as they exist for most schools (compared to other subjects in the national exams such as Swedish that is split into Swedish for beginners and regular Swedish). They are also outcomes that are continuous distributions as we want to avoid "threshold outcomes" that can lead to erroneous conclusions (Ho, 2008).

$\sum_{t=2014}^{2022} \beta_t Year_t * Inspection_i$ represents the interaction between the year variable and our treatment variable which creates a yearly *ATT* as represented by β_t. The reference year is 2013 and as the interaction requires references when $Inspection_i = 0$ the $Year_t$ serves as a time-fixed effect capturing all common time-varying variation. X_{it} represents a vector of time-varying control variables (*parents education, immigrants, boys*)[3]. ε_{it} represent a residual. Standard errors will be computed using the bootstrap procedure as per the recommendations of (Callaway & Sant'Anna, 2021). As we have no reason to believe that our sampling is clustered, we do cluster at the school and municipality level partly due to it being the assignment level (school is the assignment level but municipality could be it as well given that all schools are public schools which all are run by the municipalities) as the recommendations of Abadie et al. (2023). However, these levels are also chosen due to the panel data setting where it is potential autocorrelation within schools over time and between schools within municipalities.

Weights demonstrated in the OLS objective function:

$$min(\beta t, \beta_2) = \sum_i w_i \left(y_{it} - \sum_{t=2014}^{2022} \beta_t Year_t * Inspection_i - \beta_2 X_{it}\right)^2$$

This equation illustrates the weighting through the OLS objective function that aims to minimize the sum of squares to provide the best linear fit. Here w_i represent the IPW that weights the operation to observations that are more comparable. The IPW gives a larger weight to the "weird" observations, i.e. those that have a low propensity score but were still treated and those that have a high propensity score but are not treated. The propensity score was estimated via a logit model. Given that the schools that are inspected (observations receiving treatment) are not random but quite the opposite (they are purposefully selected based on a risk assessment) the treatment and control groups are likely to differ on a range of dimensions. Since the SSI is a government agency and bound by law to a certain transparency, the index they use to select the schools for inspection is more or less available online (The Swedish School Inspectorate, 2015, 2020a). Hence the school characteristics used for picking

[3] These are the openly available SES measures and have shown to explain educational outcomes (Swedish National Agency for Education, 2022)

out which schools that are inspected are observable and hence we can mimic this procedure to achieve a balance in baseline characteristics between the treatment and control group. However, for the difference-in-difference methodology to hold, it is not necessary to balance all covariates but rather to weight so the PTA hold. Hence the purpose of the weighting is partly to match to satisfy this assumption and partly to capture the schools overall "risk". As mentioned the SSI:s reasons for inspection is survey answers regarding school work environment and schools leadership, student educational outcomes, complaints against the school as well as relevant information from previous inspections (The Swedish School Inspectorate, 2015, 2020a). Our cross-sectional covariates used to mimic this index is the following: The schools socioeconomic composition to proxy achievement (*parents' education, immigrants, boys* all for the year 2016), number of *complaints* (constructed as number of notifications during 2015 through 2017), perceived *safety* in the school, perceived *leadership* from the principal[4], the share of *certified teachers* for the year 2016 and *students per teacher* for the year 2016). The survey items, socioeconomic variables and complaints correspond well to the SSI: s index, the *certified teachers* and *student per teacher* are chosen to capture a more general adherence to regulations which the SSI is likely to pick up on from previous inspections or collected documentation. For more details about this see Appendix 2.

Strategic and representative sampling

To make a transparent and credible analysis this section describes how we strategically trimmed the sample while still keeping it as representative as possible.

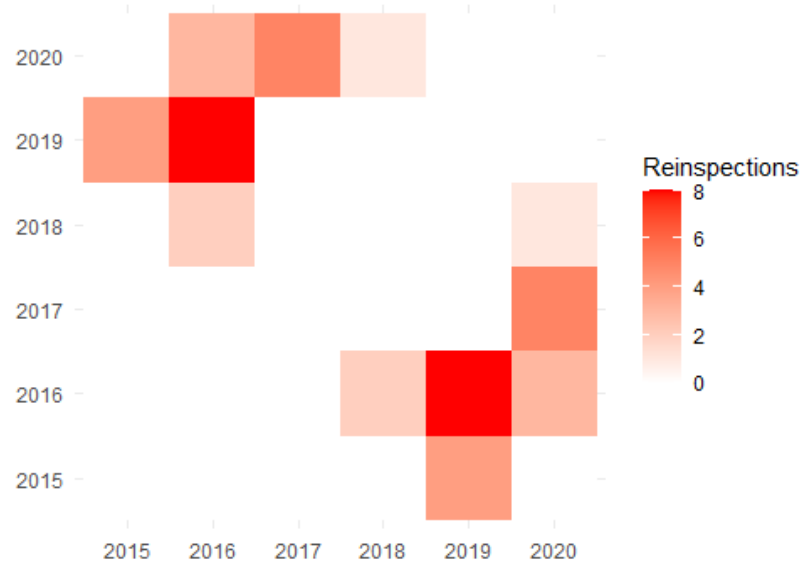
During the cycle between 2015 and summer 2018, roughly 25 % of all Swedish schools received an inspection visit. Some aspects of this cycle need to be considered when conducting this analysis, especially when it comes to implementing the double-robust difference in difference methodology. The challenges are the following: 1. The cyclical nature of the inspection visits creates schools that receive "treatment" more than once at different time points. 2. The share volume of schools inspected creates difficulty in finding roughly equally many control schools (strong overlap assumption). 3. Staggered rollout of treatment creates methodological problems (small treatment groups making it difficult to meet the

[4] For the survey variables, the lion's share of the data used has been taken from the years 2015 and 2016, but supplemented with data from 2013, 2014 and, in the last case, the spring term 2017 when data was missing.

parallel trends assumption) as well as data constraints (survey data can only be collected irregularly with only few observations prior to 2015).

When looking at how many schools received treatment "twice", Figure 2 below shows a heatmap of how many schools that got inspected in multiple years and what those years were. The heatmap clearly shows that the further from the previous inspection year the more likely a reinspection in the next cycle. Hence to avoid studying schools that were inspected twice, the sample should be limited to the years prior to the next inspection. However, that would lead to a short time window of studying the outcome of interest (maximum 3 years). Another solution to this empirical problem could be to trim the sample from schools that were inspected twice. That however would mean selecting a non-representative sample dropping 23 schools (roughly 10 % of the treated schools). A third option which we choose, is to focus on the schools inspected in the later years of the sample 2017 and during the spring 2018 (the middle to upper right corner of Figure 2).

Figure 2: Heatmap of Reinspection



Focusing on these schools inspected in 2017 and spring 2018 is beneficial for several reasons. Primarily in this group, fewer schools receive a reinspection (6 of 91 schools) which makes us more confident in trimming the sample without distorting the representativeness of it. As the COVID-19 pandemic put a halt to the inspection visits in 2020 we can exploit this exogenous

variation in inspections assuming that the schools inspected in 2017 and spring 2018 would most likely be up for reinspection as well wasn't it for the pandemic. Hence trimming the sample of the few reinspected schools in 2017 and spring 2018 leads to less distortion compared to studying all schools in the 2015 inspection cycle. Another reason for selecting these schools is due to the inspections happening close in time (all within 18 months). Due to this fact, this simplifies our research design and allows us treat it as one treatment period and not delve into the staggered treatment adoption which one risks making "forbidden comparisons" in a two-way fixed effect specification (Roth et al., 2023). Even as there are remedies to this problem by Callaway and Sant'Anna (2021) it is still preferable to avoid the a staggered design due to the difficulty meeting the PTA. As the staggard setting requires multiple treatment groups to meet parallel trends in multiple points in time, a pooling of the data over an 18-month period to a static design creates a larger treatment group at only one point in time which more easily satisfies the PTA. Trimming the sample also simplifies the design to meet the "strong overlap" assumption which is difficult to satisfy when there is an imbalance between comparison i.e. large treatment pool and small control pool. And, as the data sources used to create the "risk analysis" by the Swedish school inspection have been structurally changing over time (the SSI: s surveys to students have continuously been updated) and also are only collected irregularly fixing the treatment group to a specific point in time allows for comparable data.

When selecting the final sample several other decisions had to be taken on how to deal with missing data and schools that exited and entered during the study period. Details about this are in Appendix 1. The final sample consists of 800 public lower-secondary schools that existed in 2017 over the time period 2013-2022 where 85 of them received an inspection visit in 2017 or spring 2018. Descriptive statistics for the time-varying variables are provided in Table 1 below, split into subgroups depending on treatment status.

Table 1: Descriptive Statistics by Treatment status and Variable

Inspected Yes(N=85), No(N=715)	**Variable**	**Mean**	**SD**	**Min**	**Max**
No	Immigrants	5.188	5.829	0.00	50.00
Yes	Immigrants	7.232	7.187	0.00	38.00
No	Boys	52.132	7.618	11.00	94.00
Yes	Boys	53.000	6.902	30.00	83.00

No	Certified teachers	84.427	9.478	41.10	100.00
Yes	Certified teachers	80.238	10.447	42.30	100.00
No	Students per teacher	12.226	2.029	3.10	35.30
Yes	Students per teacher	11.636	1.845	7.10	27.20
No	English test	15.004	1.326	6.40	19.50
Yes	English test	14.327	1.244	6.70	17.90
No	Grades	224.758	22.164	130.00	301.00
Yes	Grades	212.458	20.008	146.00	278.00
No	Parents Education	2.279	0.207	1.57	2.90
Yes	Parents Education	2.161	0.198	1.54	2.84
No	Mathematics test	11.307	1.911	3.50	17.20
Yes	Mathematics test	10.372	1.899	2.60	14.70

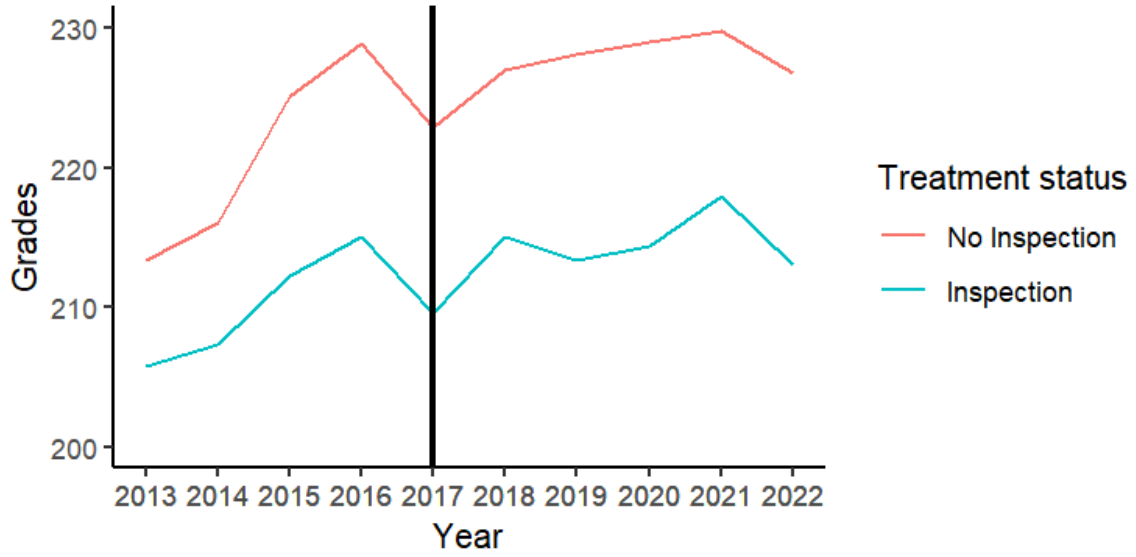
Empirical Strategy Validity

Parallel trends assumption

To assess the parallel trend assumption, we in this section present plots of the raw trends of the outcome variables in question. To this purpose, we also estimate simple regressions with time-by-treatment interactions to see if there is a statistical difference in trends prior to treatment which we also discuss in this section (see Appendix 3 for these models).

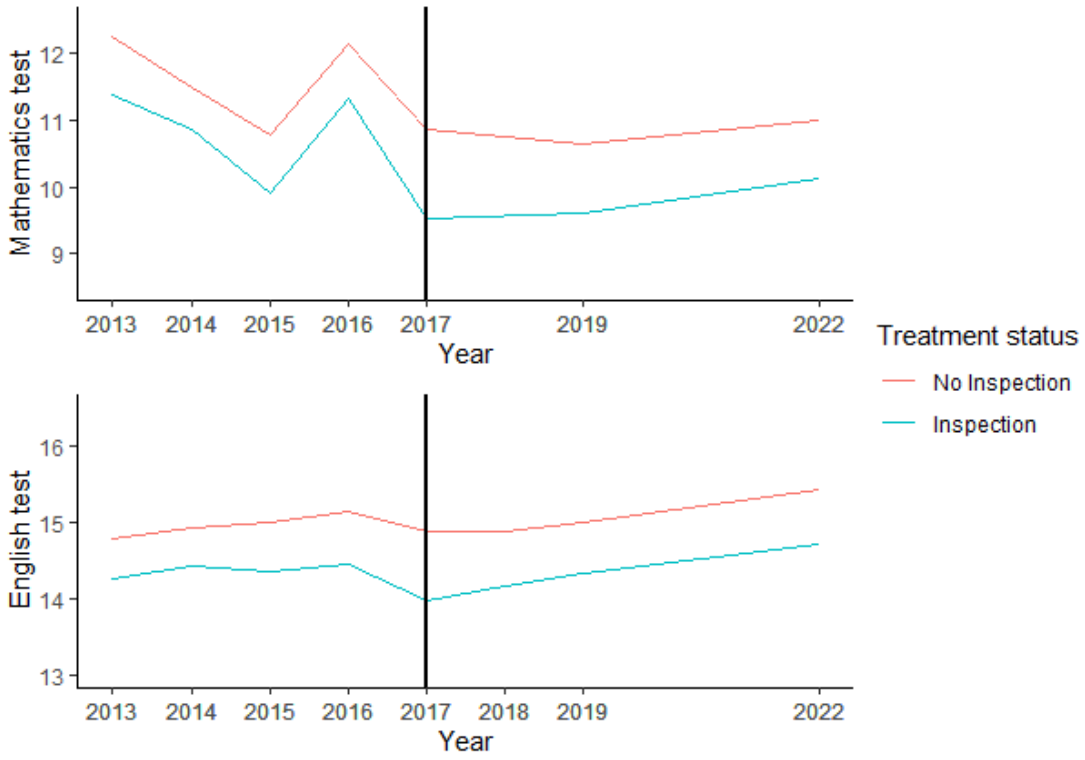
First Figure 3 displays the raw trends of the *grades*. Reading from the figure it is difficult to conclude that the trends are parallel as they roughly but not precisely follow each other over time prior to treatment. When running regressions with time by treatment interaction the year 2016 sticks out as significantly different between the two groups (see Appendix 3). Hence it is probable that the trends are not parallel prior to the intervention.

Figure 3: Raw trends of differences in Grades between inspected and not inspected schools



Next, we turn to the other outcome variables of the national tests that can be seen in Figure 4 with the *mathematics test* in the top plot and the *English test* in the bottom. Again, it is a difficult task to assess this visually, but it at least appears as there are no major deviations in the pre-treatment trends. The simple regression model also does not show any significant differences pre-treatment. In general, it appears that differences between the two groups are more profound for the grades than the national test scores indicating that either grades are simply much more complex information that differs between schools to a larger extent or that grades are a stronger indicator for the assignment of treatment.

Figure 4: Raw trends of differences in national test score (Mathematics top plot, English bottom plot) between inspected and not inspected schools



Hence based on these raw tests it appears that the violation of the parallel trends is not a severe problem, although it could potentially be for grades as the outcome variable. As already mentioned, the double-robust procedure will weigh the model to meet the parallel trends assumption conditional on the covariates in the weights as well as the control variables. This will be implemented by default in all the specifications to keep a flexible estimation framework. Given that the R command from Callaway and Sant'Anna (2021) is implemented a pre-test of the parallel trend assumption for the full model specification is provided. All models presented have a p-vale of at least 0.1 that indicates that we can't reject that the trends are parallel at the 10 % significance level. The parallel trends will also be visibly assessable in the event-study plots pre-periods.

Covariate balance

To conditionally satisfy the parallel trends assumptions it is important to examine the baseline differences in covariates between the treatment and control group. Even as the parallel trends could be met unconditionally on these covariates, or that it could be met without all the covariates necessarily being balanced it serves as a test to see that the weighting indeed creates more comparable groups. If there is a larger deviation of some of the schools characteristics it could be that these differences possibly could bias the model results. As we

would ideally balance the covariates dynamically to capture potential trends of the covariates, the data limitation only allows us to do this statically one year prior to treatment.

Table 2 below shows the baseline differences between the treatment and control groups which indicated differences in means for almost all covariates. The inspected schools have lower socioeconomic status of their students, worse educational performance, more complaints directed at them, are perceived as less safe, the principal is perceived as a lesser pedagogical leader, the school have fewer certified teachers, a higher share of boys and finally slightly surprisingly have fewer students per teacher.

Table 2: Baseline differences between treatment and control group pre-weighting.

Covariate	Mean Treated	Mean Control	Standardized mean deviation	Standardized variance ratio
Parents Education	2.161	2.279	-0.581	0.910
Immigrants	7.232	5.188	0.312	1.520
Boys	53	52.132	0.119	0.821
Complaints	1.288	0.814	0.328	2.278
Safety	7.842	8.015	-0.313	1.399
Leadership	7.008	7.359	-0.333	1.146
Students per teacher	11.636	12.226	-0.304	0.827
Certified teachers	80.238	84.427	-0.420	1.215
Mathematics test	10.372	11.307	-0.491	0.988
English test	14.327	15.004	-0.527	0.880
Grades	212.458	224.758	-0.583	0.815

In Table 3 below we display the differences post-weighting. Most of the covariates are now balanced on the mean (smd <0.1). A few exceptions are *mathematics test* and *English test*, *boys* and *leadership* that are just above the rule-of-thumb of 0.1. we don't see this as a large problem as they are compared to previously much closer in means and both their variances are now more balanced between the control and treatment groups (Standardized mean variance is close to 1). The large exception is the data on perceived *safety* where the group with higher perceived safety now has flipped to the treatment group and is still different in mean and variance. This does not mean that the weighting has failed in its intended purpose but rather it results in a compensation too much for this covariate. As most other covariates are balanced, we perceive this as a necessary imperfection.

Table 3: Baseline differences between treatment and control group post-weighting.

Covariate	Mean Treated	Mean Control	Standardized mean deviation	Standardized variance ratio
Parents Education	2.282	2.268	0.062	1.312
Immigrants	5.602	5.388	0.033	1.265
Boys	53.022	52.166	0.117	0.857
Complaints	0.906	0.867	0.097	0.881
Safety	8.115	7.999	0.599	1.574
Leadership	7.381	7.323	0.159	1.161
Students per teacher	12.130	12.186	-0.030	0.818
Certified teachers	83.630	84.072	-0.049	1.020
Mathematics test	11.029	11.239	-0.128	1.084
English test	14.821	14.952	-0.112	0.979
Grades	222.670	223.886	-0.054	1.055

Bad controls

When utilizing time-varying covariates in the outcome regression as controls it is important to check for the "bad control problem" (Roth et al., 2023). If our time-varying *covariates parents' education, immigrants* and *boys* are affected by the treatment this could induce bias. Hence, we test for this by entering our covariates as outcome variables in the doubly robust difference-in-difference to see if they are impacted by the treatment. We find no statistically significant effect of the treatment on these variables. For details see Appendix 4.

Anticipation

The difference-in-difference model require the assumption of no anticipation before treatment (Roth et al., 2023). Schools at this point of time got notified roughly three weeks prior to inspection meaning that potential anticipation from notification seems impossible one year

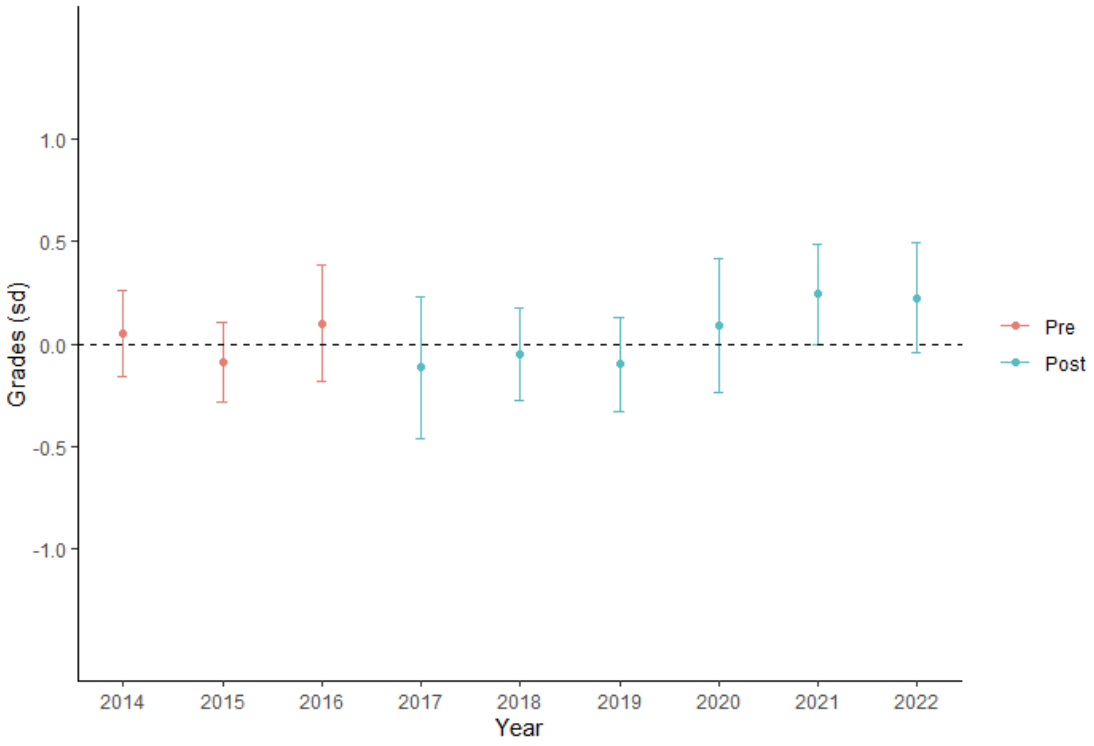
prior in the data. However, schools could potentially be aware that a certain educational performance can attract the attention of the SSI. That is a built-in mechanism that the "threat" of inspection could get the schools to act but given that this this "threat" is present for all schools we see it as unlikely that the treatment group would anticipate this to a greater extent as the control group as they are similar in levels of grades as seen in Table 3.

5. RESULTS

In this section, I will display the results of the main specification with doubly robust difference-in-difference for the different outcome variables following the R command from (Callaway & Sant'Anna, 2021). All educational output variables are scaled and expressed as standard deviations. The full model results are available in Appendix 5.

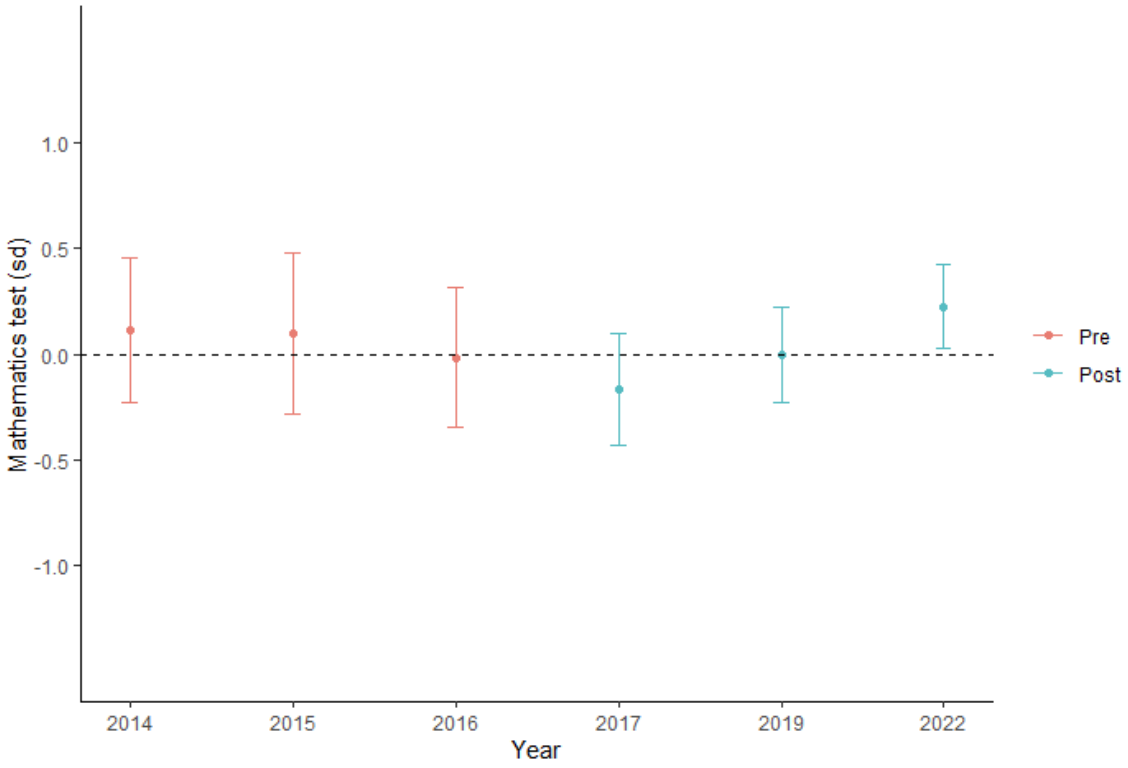
Regarding the model with *grades* as the outcome variable Figure 5, there is a plot that shows the dynamic treatment effect over the years. Each dot represents the treatment by year interaction and the error bar shows a 95 % uniform confidence band. The full specification indicates a statistically significant effect for the year 2021 that corresponds to 0.24 standard deviation.

Figure 5: Dynamic effect of being Inspected on Grades



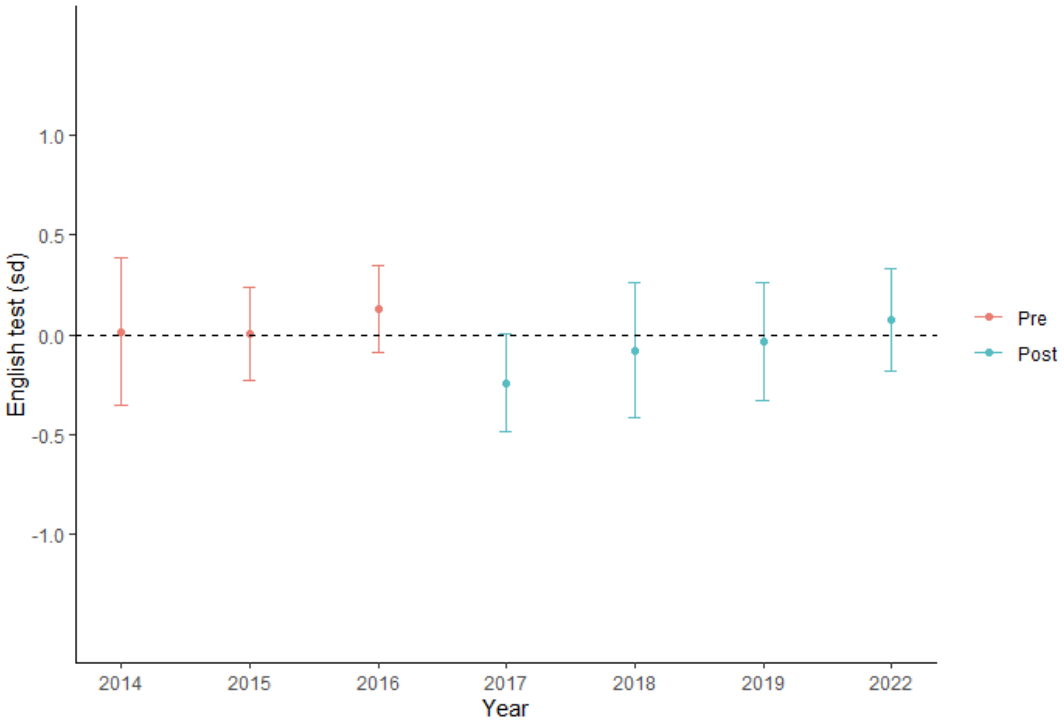
The model of the national *mathematics test* is Figure 6 below. Unlike grades there appears as with the mathematics test as an outcome variable there is a statistically significant effect for the year 2022 i.e. roughly 5 years after the inspection. The effect corresponds to 0.23 of a standard deviation.

Figure 6: Dynamic effect of being Inspected on National test score in Mathematics



Finally, we examine the national *English test* which is displayed in Figure 7 below. In this model, there are no statistically significant effects. However, the estimate for the main treatment year 2017 is close to being negatively significant.

Figure 7: Dynamic effect of being Inspected on National test score in English



Balancing the panel

As a robustness test, we balance the panel to omit schools that enter and exit the dataset over time. For the treatment group, it is roughly 14 schools that exit before 2022. After balancing the panel there are only 55 treated schools left which could be too few to find a treatment effect. When running the models (available in Appendix 6). The results are similar for *grades* and *English test* but for the *mathematics test,* the positive estimate in 2022 is no longer significant and slightly lower in magnitude. This shows that it is important to consider the entry and exit of schools over time, but it remains unclear if it is due to the small treatment group or the balancing itself that changes these results. Overall, the positive "trend" of estimates remains although not statistically significant.

Placebo test

I also conduct a placebo test to see if the results change when changing the treatment to a random sample within the "worst quartile" of the schools according to their propensity scores. Below is a balance table showing the means of the covariates for the placebo group compared

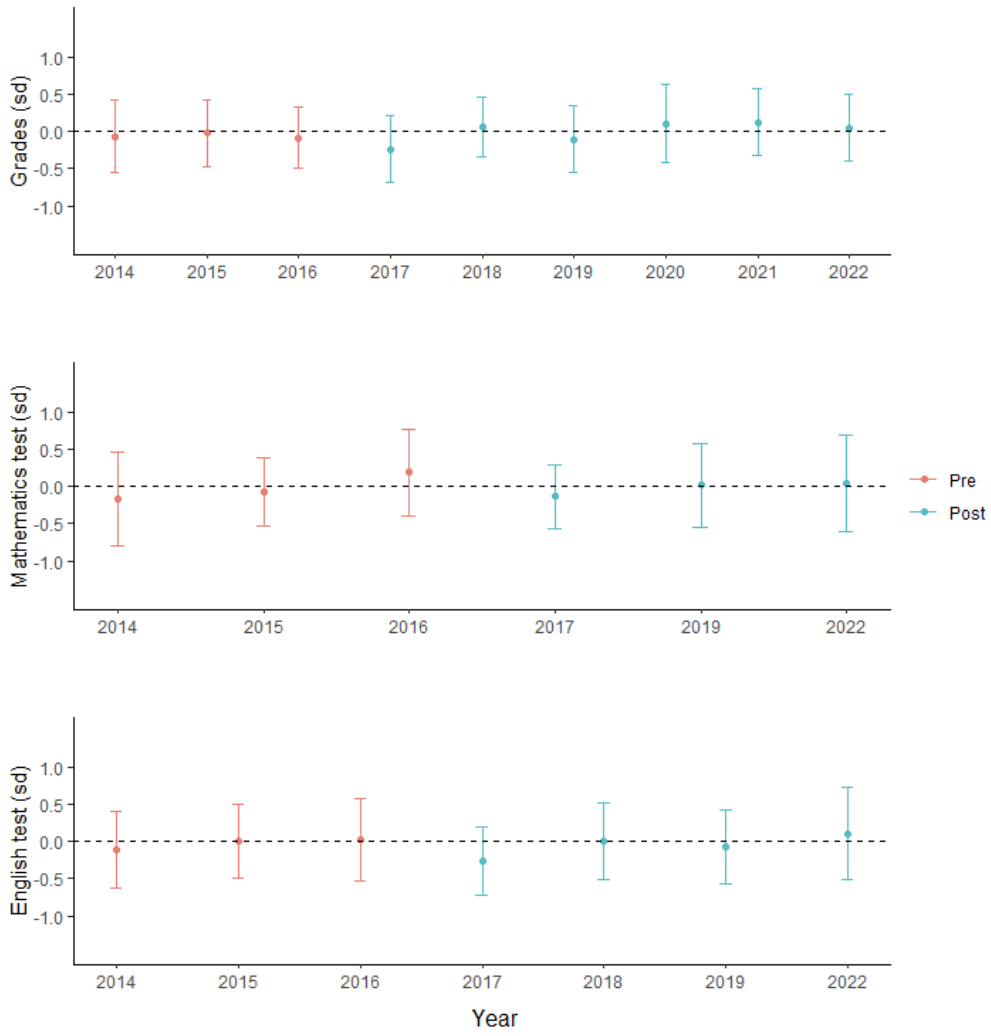
to the means of the real treatment group and control group. Their means are roughly similar to the real treatment group.

Table 4: Means of control, treatment and placebo group

	Immigrants	Parents Education	Boys	Grades	Mathematics test	English test	Certified Teachers	Students per teacher
Control	4.864583	2.299313	52.09763	226.7466	11.44449	15.11583	85.14216	12.34316
Placebo	8.314246	2.112416	52.72626	209.6020	10.16444	14.08566	78.09483	11.30626
Treated	7.418206	2.156359	52.92480	211.9129	10.30172	14.27626	80.01009	11.57259

Figure 8 below shows the result of the placebo test. It appears that none of the effects in the real analysis shows up in this placebo test. Thus, it is probable that the effects found for the full analysis are real and not spurious.

Figure 8: Placebo test with conditionally random sample of inspected schools



Changing the control group

As another robustness test the control group the control group is changed. Given that the regressions previously contain all schools (although weighted) it is interesting to see what happens when we remove the schools with lowest propensity to be inspected. This to see if we can get closer to a "apples to apples" comparison. Thus, in the first test 25 % of the control group is removed, a point of which roughly 90 % of the inspected schools are above in propensity to be inspected. This can be seen in Figure 9. Then in the second test half of the control group is removed below the median propensity score, a point of which 75 % of the inspected schools are above in propensity to be inspected. This can be seen in Figure 10.

The figures reveal that the result for the *grades* is no longer significant but given how close it is it is difficult to say that its disproven. Otherwise, the results for the *mathematics test* model are robust to this test (point estimate even slightly stronger in magnitude) while for the *English test* negative effect for the year 2017 is now significant with a negative point estimate around 0.26-0.3 standard deviation.

Figure 9: Trimmed control group test with 25 % lowest propensity schools removed

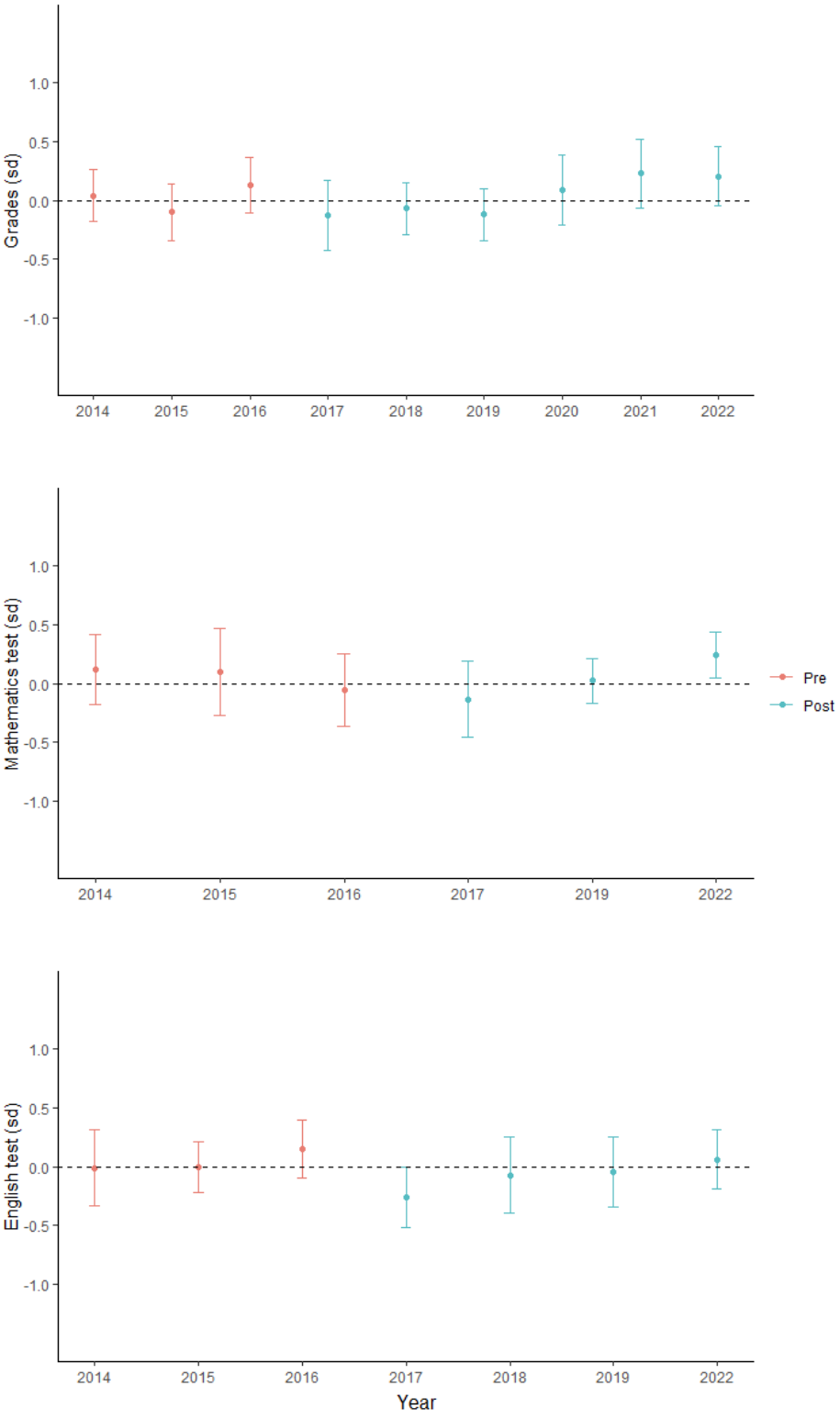
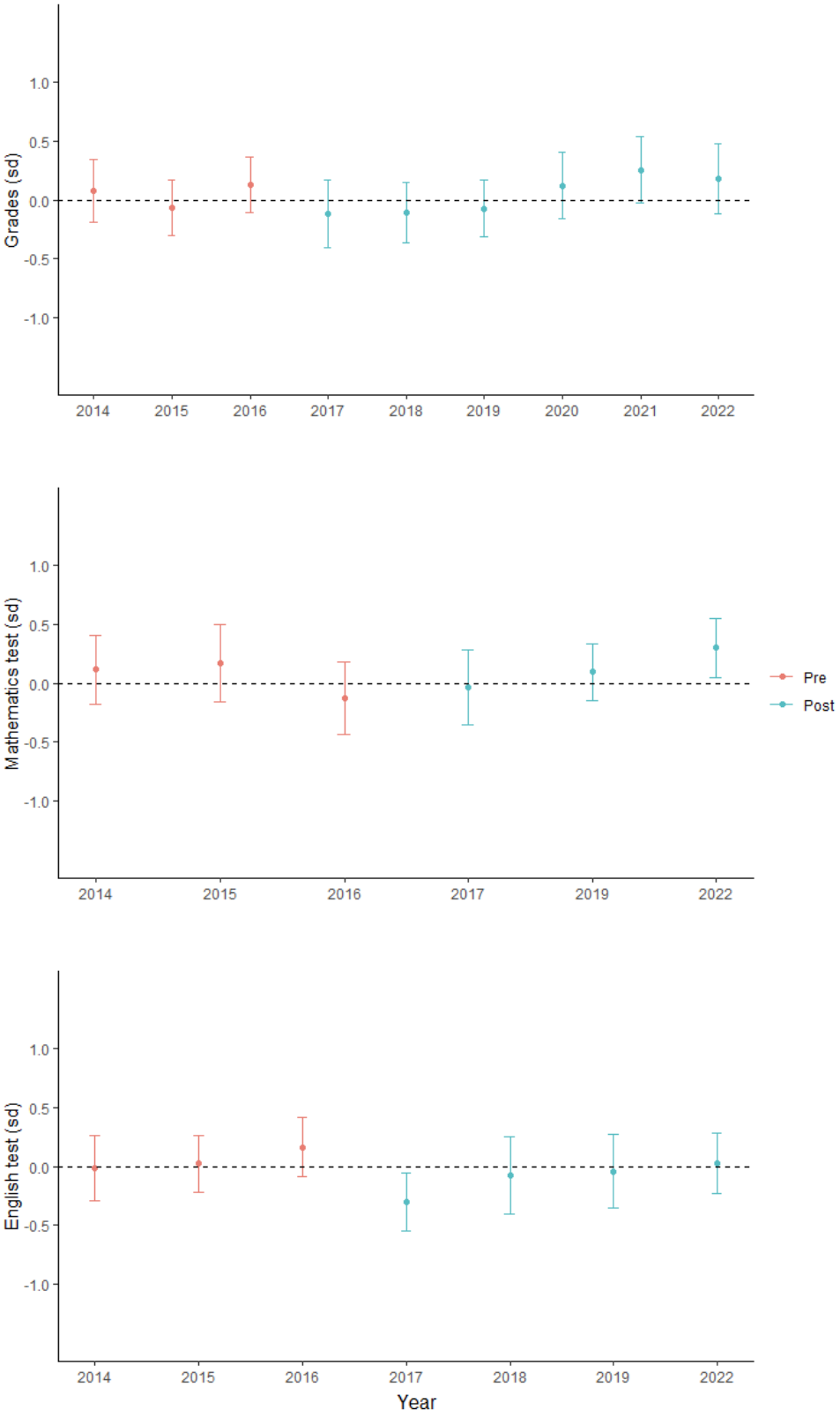

Grades (sd)
1.0
0.5
0.0
-0.5
-1.0
2014
2015
2016
2017
2018
2019
2020
2021
2022
Mathematics test (sd)
1.0
0.5
0.0
-0.5
-1.0
2014
2015
2016
2017
2019
2022
Pre
Post
English test (sd)
1.0
0.5
0.0
-0.5
-1.0
2014
2015
2016
2017
2018
2019
2022
Year

Figure 10: Trimmed control group test with 50 % lowest propensity schools removed



Allowing for two time periods

As a test, we run the command with the treatment group split into its respective years 2017 and 2018. We do this since pooling them in one time period (18 months) could be a potentially erroneous simplification if there is heterogeneity in treatment effects across time. Hence the Callaway and Sant'Anna (2021) estimator allows for staggered treatment adoptions over time.

Hence the 85 treated schools are split into two groups for 2017 with 57 of the treated schools and one for 2018 with the remaining 27. A difficulty with splitting the sample is that the treatment groups become very small, especially for the year 2018. Hence this results in less stable trends over time which makes it more difficult to meet the PTA. As a matter of fact, for the model with *grades* as an outcome variable, we do not seem to be able to meet it as the pre-test rejects the null hypothesis of parallel trends. However, we do present the results for the national exam in *mathematics test* (Figure 11) and *English test* (Figure 12) below where the pre-test holds for both models. The model for the Math test now shows a statistically significant negative effect for the year 2017 for the treatment group that is treated 2017 which corresponds to a 0.3 standard deviation lower score in Mathematics the year during inspection. The negative effect for the year 2017 for mathematics corresponds well to the intuition that the inspection visit can be disruptive since the school spend time and resources on "window dressing" (Alvesson & Strannegård, 2021; Rosenthal, 2004). The positive effect for the year 2022 is no longer statistically significant in this model. For the English model, we now find a statistically significant effect of 0.31 standard deviation higher score for the year 2022.

Figure 11: Mathematics test split into two time periods

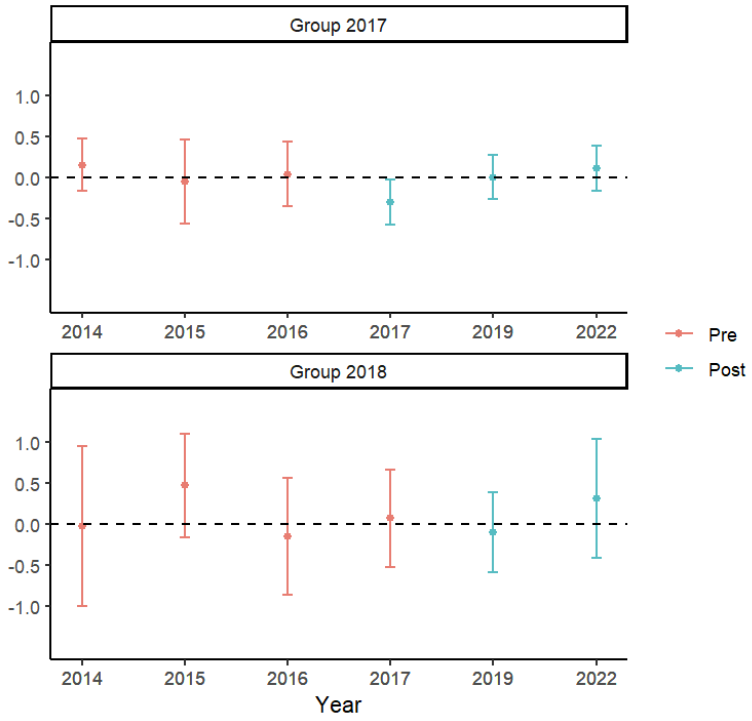
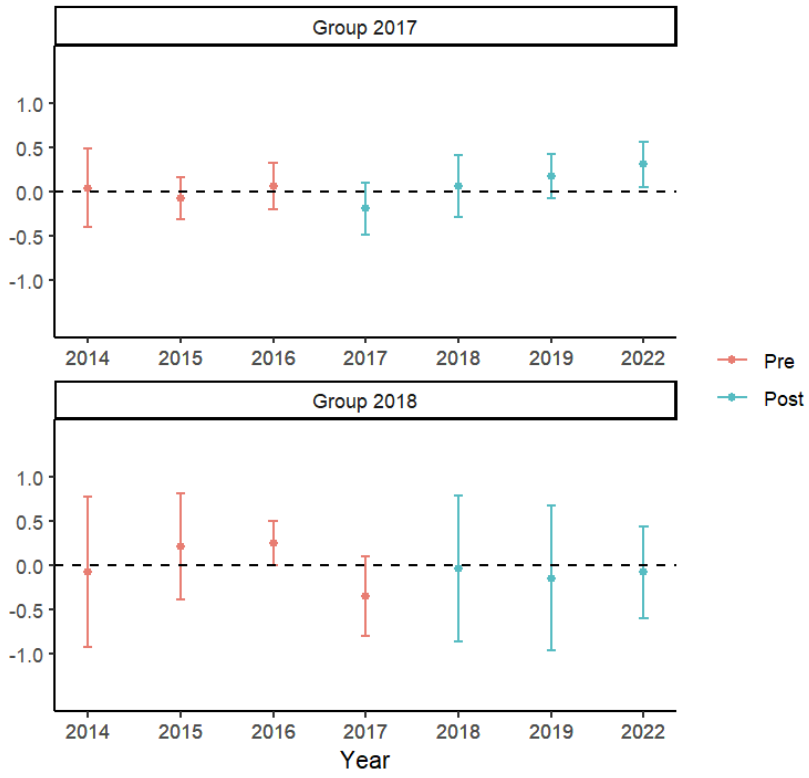


Figure 12: English test split into two time periods



These results imply indeed that it is important to treat the time periods as separate. However, given the data limitations in this study, this has to be considered a trade-off between keeping a large enough treatment group to study this at all, while still not manipulating the actual treatment periods.

Heterogeneity analysis

As a heterogeneity analysis, I do a split sample analysis where I break the sample into two based on the two dimensions of the school compositions. I look at the effect of inspection of schools with a higher and lower share of students whose parents have a certain level of educational attainment as well as schools that have a higher or lower share of immigrants. The sample is split to achieve a balance of treated observations in both groups without distorting the sample sizes to any larger degree. The rule-of-thumb I use is that I split as close to the median as possible while still keeping at least 40 % of treated observations in one of the groups. For the parent's education that results in me cutting at the 30^{th} percentile and for immigrants at the 65^{th} percentile. I guarantee the sample size to still be more than 2000 observations in each. The dimensions of heterogeneity where chosen partly due to previous studies mostly has investigated heterogeneity by different socioeconomic measures (income, education performance, ethnicity).

In Figures 13 and 14 below I display the differences in effects depending on the share of students whose parents have a certain level of educational attainment (Figure 13) and the share of students who are immigrants (Figure 14). Since some models fail the pre-test of parallel trends for heterogeneity by parents' educational attainment, I only display the results for *Mathematics test* since it is the only one passing the pre-test. For the share of immigrants, the *Grades* and *Mathematics test* model pass the pre-trend test.

We see only small and non-significant differences in effect depending on parents education in the school. This does not seem to correspond to the findings of Figlio and Loeb (2011); Hussain (2015) which finds stronger effects for "weaker schools". We see no effect depending on the immigrant status either. As shown in previous literature minority group status do not appear to be an crucial moderator (Figlio & Loeb, 2011).

Figure 13: Heterogeneity of effects depending on students' parents' average educational attainment

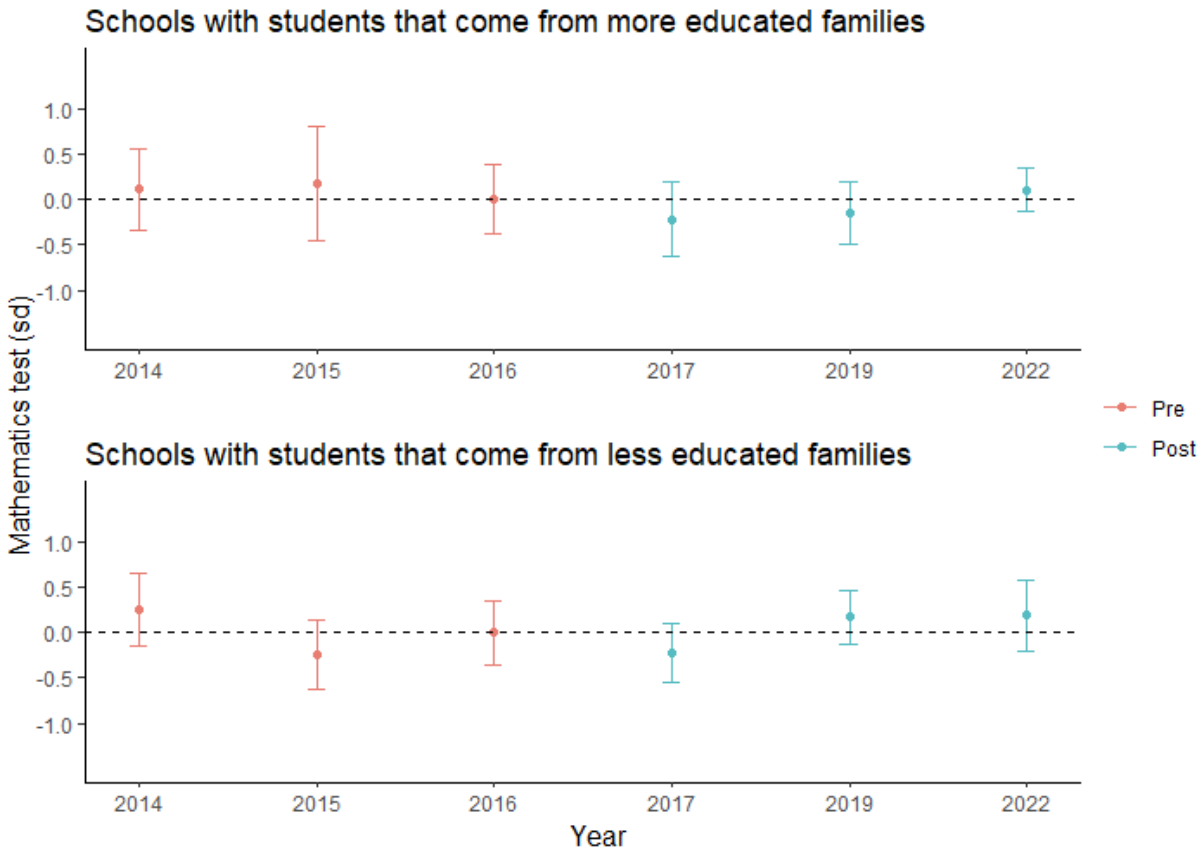
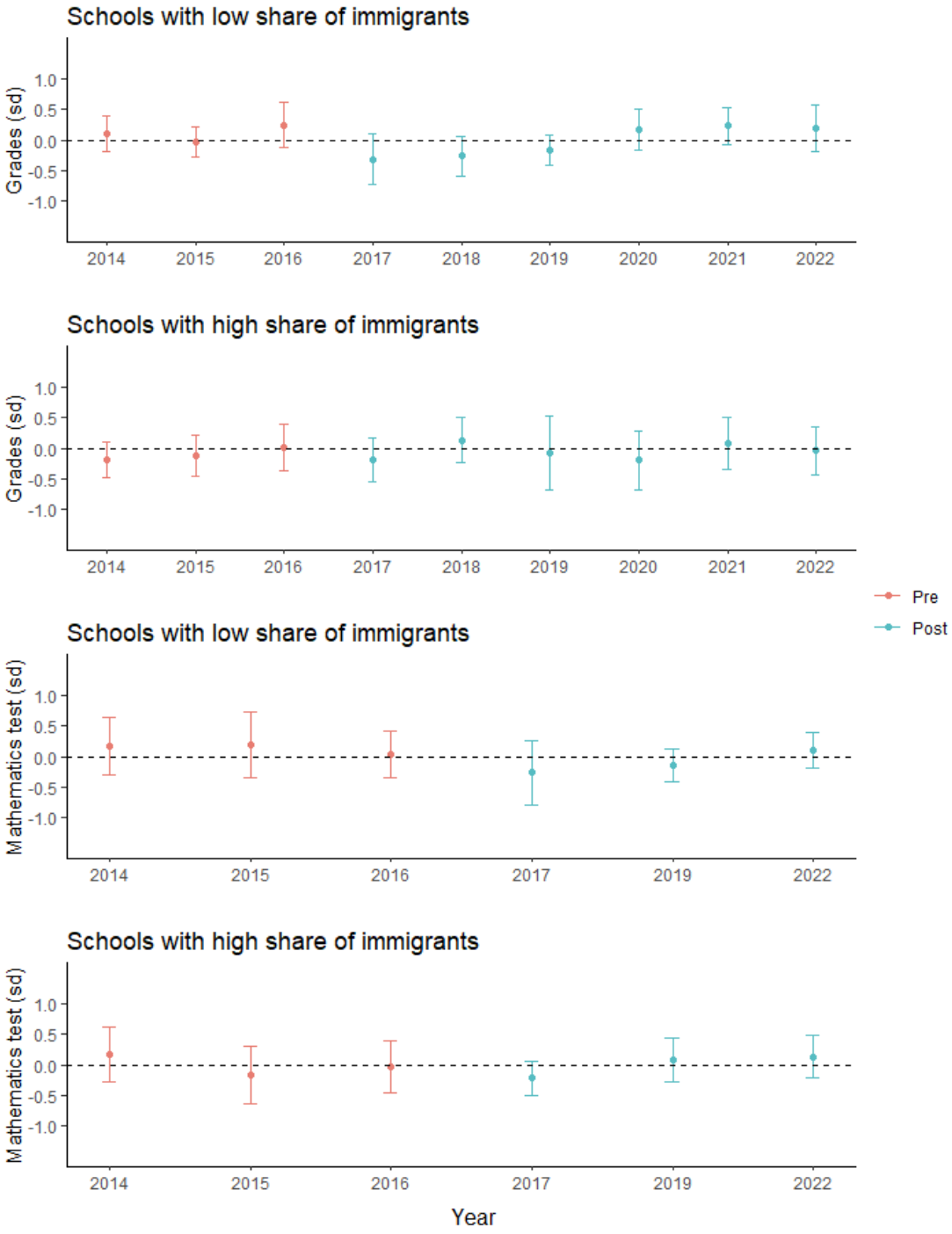


Figure 14: Heterogeneity of effects depending on share of immigrants

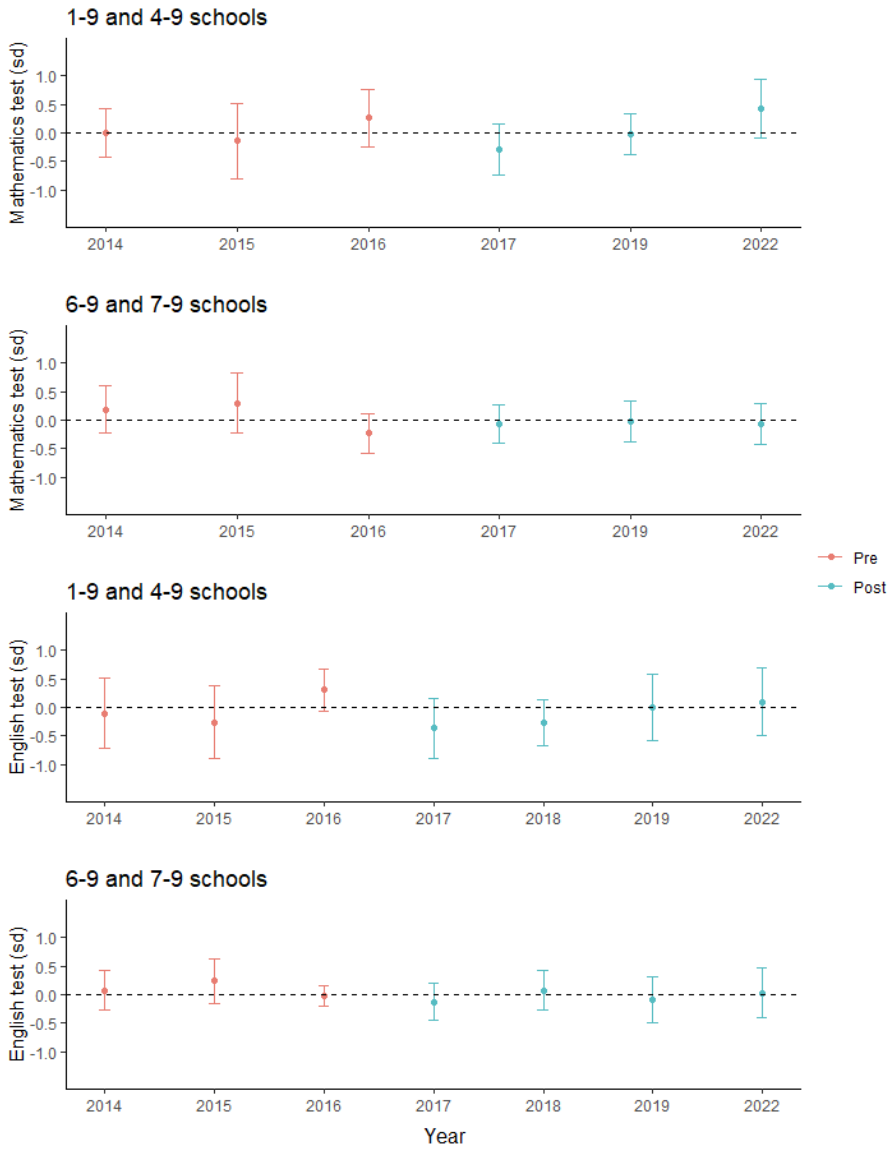


Heterogeneity of school structure

As a final heterogeneity analysis, we split the sample by the structure of the school. That is if the school covers at least 5 grade cohorts (1-9 and 4-9 schools, i.e. a 9 to 5 year school) or less than 5 grade cohorts (6-9 or 7-9 schools i.e. a 4 to 3-year school). This is done to examine our longer-term effect for the year 2022. If the effect is true and not spurious it can be partly driven by how long the students in the schools attend the treated schools. If it has the structure of being at least 5-grade cohort school i.e. a 9 to 5 year school the students exiting the 9th

grade in the year 2022 could be attending the school from earliest the 2013 and through the inspection years 2017 or 2018. If the school is less than a 5-grade cohort school i.e. a 4 to 3 year school the students exiting the 9th grade in the year 2022 would have started the school by earliest the autumn of 2018 and forwards. Hence the second group would not be "treated" for as long and we would expect smaller effects for this one. In Figure 15 below we show the results for the *mathematics test* and *English test*. The results revealed some heterogeneity for the mathematics test. The schools with more grade cohorts have a substantially larger point estimate than the ones with fewer grade cohorts. Although it is not statistically significant so it should be interpreted carefully.

Figure 15: Heterogeneity of effects depending on school structure



6. DISSCUSSION

To untangle these results, it is important to keep in mind that different treatment group sizes and treatment group compositions such as the balanced panel test, multiple period tests and heterogeneity analysis need to be considered. A difference in results could be due to the sizes of treatment groups. For example, the balanced panel test in Appendix 6 does not necessarily disprove the results found for the test in mathematics in the full model, it could be due to lower treatment group size or of course, the unbalanced panel in the full model could show us spurious results. However, given that the placebo test shows no effect in the mathematics model the place test does show that the problem with the declining treatment group does not replicate in this setting even though it has similar problems in terms of an unbalanced panel. Hence, it could be argued that the placebo test proves the point that the results of the mathematics test are not spurious. It is however more difficult to know for the split time-period analysis if the results are due to smaller treatment groups or not. Thus, all results should be carefully interpreted.

First off, given that there is a consistent effect on grades in both the full model and the balanced panel test for the year 2021 (while not showing up in the placebo test). We can conclude that inspection visits increase grades by roughly 0.2-0.25 standard deviation 4 years post the main inspection year. Why does this effect show up for this year and not the others? It could be the case that the treatment and control groups have different patterns in terms of grade inflation that could be seen this year due to the cancellation of the national tests that usually are the benchmark for grade inflation. Hence if inspected schools were more incentivized to "prove themselves" they could be more likely to inflate grades when given the opportunity as during 2021.

Regarding the mathematics test that is more robust to grade inflation given the structure of the examination and subject matter, there is less reason to consider any inflation mechanism. The results found for the year 2022 are interesting and could be an indication that schools are incentivized by inspection and are successful at improving roughly 5 years post-inspection. Given that previous literature shows that school accountability is most effective at impacting mathematics this is plausible (Figlio & Ladd, 2014; Hofer et al., 2020; Luginbuhl et al., 2009). However, this result does not replicate under the split-sample analysis which could be due to smaller treatment groups, but also the original model could be erroneous due to pooling the time-periods. Since it is substantially heterogeneity in effects regarding on the school structure (larger point estimate for schools whose students are "treated" over a longer period)

it is also an indicating that the estimate is true and that the schools are improving over time. If the treatment effect is true and is a delayed improvement of 0.23 standard deviation this calls for optimism. For now, it should suffice to conclude that school accountability needs to be studied long-term to detect effects like this.

In the split-sample two time periods analysis, a negative effect for the mathematics tests in the year of inspection (for the 2017 group) is also found corresponding to a 0.3 standard deviation. This effect does show up for English test as well in the control group test it with a negative significant effect around 0.26-0.3. However, given that the other models deal with pooled data over time, it would not be surprising that this effect is stronger when the treatment group is only over a 12-month period rather than an 18-month period in conjunction with the focal treatment year.

For the English test, it shows no significant effects in its baseline but does so in the split-sample analysis with a positive significant effect for in the year 2022 of 0.31. This indicates that when allowing for two time periods the full dynamics over time is captured compared to the pooled model. This could be due to the later treated observations (spring 2018) not having an effect in their relatively shorter term. However, more evidence is needed to fully make any conclusive statements here regarding the English national test.

In terms of heterogeneity, it is surprising to find no effects across the examined dimensions (except school structure), especially the parents education given previous literature (Figlio & Loeb, 2011; Hussain, 2015). It could be the case that a more equitable school system such as the Swedish one in terms of resources (Holmlund et al., 2020) neutralizes any heterogeneity that was previously found primarily in a U.S setting. The null effect depending on student national background is perhaps less surprising given the mixed results in the literature(Figlio & Loeb, 2011; Hanushek & Raymond, 2005). In sum, the benefits or adverse consequences of school inspections do not seem to be moderated by the differences between these groups. This implies that the distributional consequences of this policy measure are neutral.

Finally, it appears that school inspections could matter for educational outcomes. Even as the results are not resounding evidence of either positive or negative effects of school inspection, they reveal a heterogeneity over time that needs to be considered in future studies. This emphasizes the need to study this type of policy in the longer term and studies only looking at short time windows may reach erroneous conclusions. If these indicative results are true, it does show a potential trade-off between current students and future students' educational

progress. Hence, more research on this is needed to weed out if this is the case, since then school inspections may have genuine distributional consequences over time.

7. CONCLUSION

This study has examined the effect of school inspections on educational outcomes utilizing state-of-the-art econometric techniques that are robust to misspecification. The study contributes to the literature by revealing the dynamic effect of school inspections over a longer time horizon than previously studied. The application is using Swedish data examining schools inspected in 2017 and spring 2018 and find generally negligible net effect of school inspection. There is a robust positive effect on grades for the year 2021, which could be due to increased grading leniency for that year due to the abolishment of the grade benchmark (the national exams) for that year due to Covid-19 pandemic. The results also indicate a longer-term effect of test scores in mathematics 5 years post-inspection while there is also an indication of short-term negative effect of test scores the year of inspection. As the results are somewhat contradictory and unstable in the analysis it is difficult to draw any strong conclusions from this. However, the indicative results demonstrate heterogeneity over time. The results reveal a potential trade-off between future and current students' educational outcomes which highlight the dynamically complex effects of school inspection.

This study comes with several limitations. First, the study uses a specific sample at a specific point in time, which does not necessarily make the findings generalizable. The study is also conducted at the school level of analysis which is the level treatment is assigned at but risks masking potential micro-level effects and falling into the "ecological fallacy". The study does not capture the potential pre-emptive effects of school inspection. It could be the case that the mere existence of the school inspection improves schools.

This study suggests that policymakers need to seriously evaluate different policy measures to fully understand their consequences. The Swedish School Inspection has swayed from outsourcing an evaluation of the agency's inspections against hard educational outcomes. This, despite being one of their core goals. Hence, in the future policymakers should make sure that the accountability system works from the top-down. If the inspection agencies don't want to improve through evaluation they might lose legitimacy.

Future research should continuously study these phenomena with longer dynamic effects in mind. It is important to fully understand these effects in the short and long run to make reasonable policy decisions. Future research should also explore potential consequences in terms of the schools' resources, e.g. how does the inspection impact the staff?

REFERENCES

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, *138*(1), 1-35.

Alvesson, M., & Strannegård, M. (2021). *Check, check, check-Skolinspektionen och granskandets vedermödor*. Studentlitteratur AB.

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of econometrics*, *225*(2), 200-230.

De Wolf, I. F., & Janssens, F. J. (2007). Effects and side effects of inspections and accountability in education: an overview of empirical studies. *Oxford Review of education*, *33*(3), 379-396.

Ehren, M. C., Gustafsson, J.-E., Altrichter, H., Skedsmo, G., Kemethofer, D., & Huber, S. G. (2015). Comparing effects and side effects of different school inspection systems across Europe. *Comparative education*, *51*(3), 375-400.

Figlio, D., & Loeb, S. (2011). School accountability. *Handbook of the Economics of Education*, *3*, 383-421.

Figlio, D. N., & Ladd, H. F. (2014). School accountability and student achievement. In *Handbook of research in education finance and policy* (pp. 194-210). Routledge.

Gustafsson, J.-E. (2014). *Impact of school inspections on teaching and learning in primary and secondary education in Sweden*.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, *24*(2), 297-327.

Hayes, W. (2008). *No child left behind: Past, present, and future*. R&L Education.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational researcher*, *37*(6), 351-360.

Hofer, S. I., Holzberger, D., & Reiss, K. (2020). Evaluating school inspection effectiveness: A systematic research synthesis on 30 years of international research. *Studies in Educational Evaluation*, *65*, 100864.

Holmlund, H., Sjögren, A., & Öckert, B. (2020). *Jämlikhet i möjligheter och utfall i den svenska skolan*.

Hussain, I. (2015). Subjective performance evaluation in the public sector: Evidence from school inspections. *Journal of Human Resources*, *50*(1), 189-221.

Jacob, B. (2017). The changing federal role in school accountability. *Journal of Policy Analysis and Management*, *36*(2), 469-477.

Luginbuhl, R., Webbink, D., & De Wolf, I. (2009). Do inspections improve primary school performance? *Educational evaluation and policy analysis*, *31*(3), 221-237.

McElroy, K. (2023). Does test-based accountability improve more than just test scores? *Economics of Education Review*, *94*, 102381.

Meyers, C. V., & Smylie, M. A. (2017). Five myths of school turnaround policy and practice. *Leadership and Policy in Schools*, *16*(3), 502-523.

Murphy, J., & Meyers, C. V. (2007). *Turning around failing schools: Leadership lessons from the organizational sciences*. Corwin Press.

Ramböll. (2023). Skolinspektionen och kvalitet i Skolan. In.

Rosenthal, L. (2004). Do school inspections improve school quality? Ofsted inspections and school examination results in the UK. *Economics of Education Review*, *23*(2), 143-151.

Roth, J., Sant'Anna, P. H., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of econometrics*, *235*(2), 2218-2244.

Sant'Anna, P. H., & Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of econometrics*, *219*(1), 101-122.

SOU. (2007:101). Tydlig och öppen. Förslag till en stärkt skolinspektion. In: Fritzes Stockholm.

Strategirådet. (2016). *Hur fungerar regelbunden tillsyn? – Utvärdering av Skolinspektionens modell för regelbunden tillsyn*.

SVT. (2018). *Nationella prov har spridits på sociala medier*. https://www.svt.se/nyheter/inrikes/nationella-prov-har-spridits-pa-sociala-medier

Swedish National Agency for Education. (2022). *SALSA - en statistik modell*. https://www.skolverket.se/skolutveckling/statistik/salsa-statistisk-modell
The Swedish School Inspectorate. (2015). *Annual thematical report 2015*. Retrieved from https://www.skolinspektionen.se/beslut-rapporter/publikationer/regeringsrapporter/2016/arsrapport-2015-okat-fokus-pa-skolor-med-storre-utmaningar/
The Swedish School Inspectorate. (2017). *Annual report 2017*. Retrieved from https://www.skolinspektionen.se/beslut-rapporter-statistik/publikationer/regeringsrapporter/2018/arsredovisning-2017/
The Swedish School Inspectorate. (2018). *Annual report 2018*. Retrieved from https://www.skolinspektionen.se/beslut-rapporter-statistik/publikationer/regeringsrapporter/2019/arsredovisning-2018/
The Swedish School Inspectorate. (2019). *Annual report 2019*. Retrieved from https://www.skolinspektionen.se/beslut-rapporter-statistik/publikationer/regeringsrapporter/2020/arsredovisning-2019/
The Swedish School Inspectorate. (2020a). *Anledning till Inspektion*. https://www.skolinspektionen.se/inspektion/inspektion-steg-for-steg/infor-inspektion/urval-for-inspektion/
The Swedish School Inspectorate. (2020b). *Uppdrag och Verksamhet*. https://www.skolinspektionen.se/om-oss/uppdrag-och-verksamhet/
The Swedish School Inspectorate. (2023). *Annual report, 2023*. Retrieved from https://www.skolinspektionen.se/beslut-rapporter/publikationer/regeringsrapporter/2024/arsrapport-2023/
The Swedish School Inspectorate. (2024a). *Inspektion-steg för steg*. https://www.skolinspektionen.se/inspektion/inspektion-steg-for-steg/under-inspektion/
The Swedish School Inspectorate. (2024b). *Så bedöms skolor*. https://www.skolinspektionen.se/inspektion/inspektion-steg-for-steg/under-inspektion/sa-bedoms-skolor/

APPENDIX 1

The sample for this study is collected from the database of the Swedish National Agency for Education. The final sample size is of 800 public lower-secondary schools in Sweden. In this part of the appendix, I will give a more extensive description of how the final sample was constructed.

The first condition I used when sampling the schools was that they had to exist in the year 2017. Given that this is our main treatment year, it is a reasonable assumption that they did not enter post this year of exit prior to this year as we follow from 2013 to 2022. I found it reasonable to ensure that it was likely to follow the schools over time around the treatment period of interest. For the year 2017, roughly 1423 lower secondary schools existed in the dataset private and public. I omitted the independent private school due to them following a different inspection logic (all are inspected within the cycle). Prior to this omission, I used multiple imputation techniques to impute the variables of *safety* and *leadership* that came from the school surveys. Hence for the imputation, I keep the independent schools to utilize as much information as possible. The missing data for these two variables does not appear to be systematically missing when comparing the pattern of missingness to our three time-varying covariates (See red and blue bar on y-axis of plots).

Figure A1: Missing data test for Leadership against Parents Education

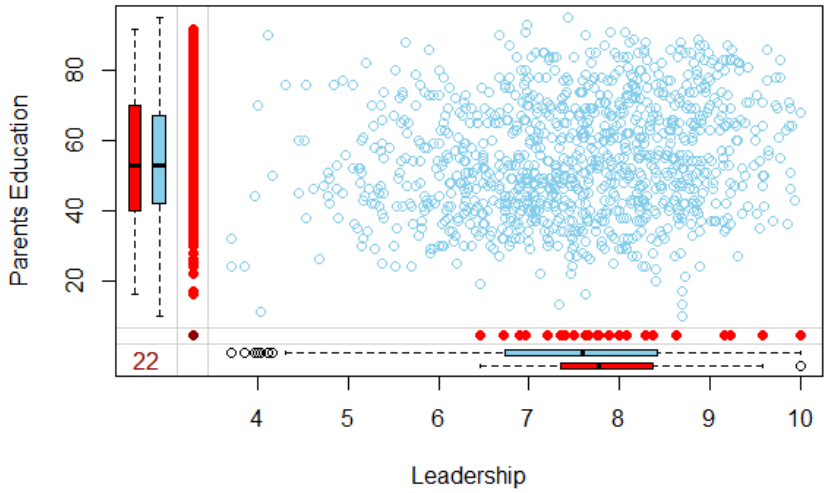


Figure A2: Missing data test for Leadership against Immigrants

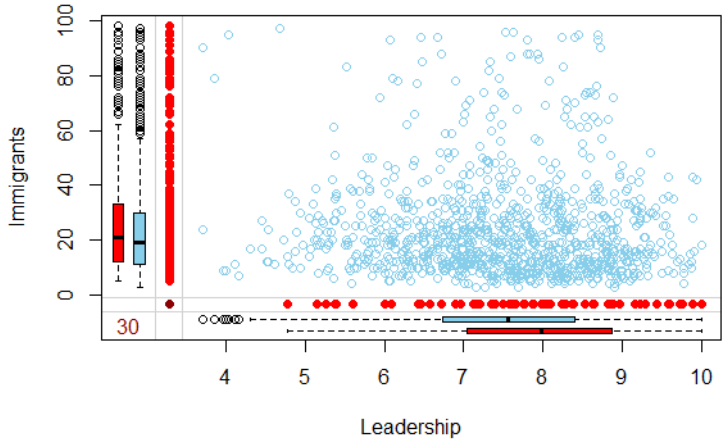


Figure A3: Missing data test for Leadership against Boys

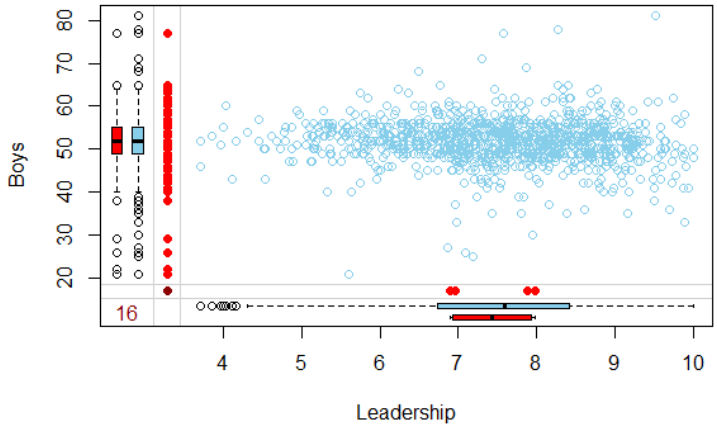


Figure A4: Missing data test for Safety against Parents Education

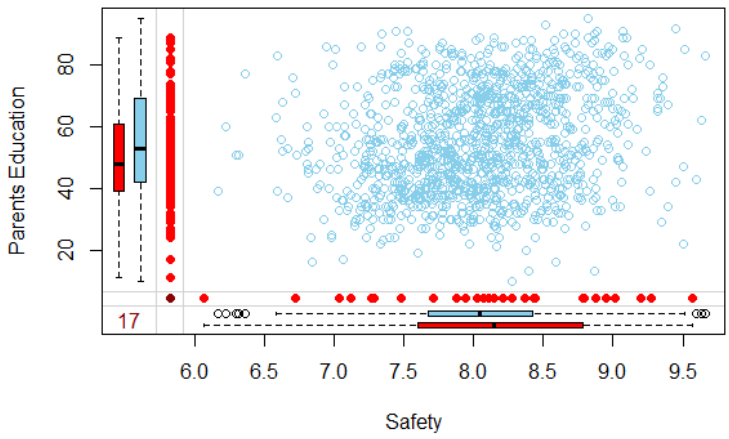


Figure A5: Missing data test for Safety against Immigrants

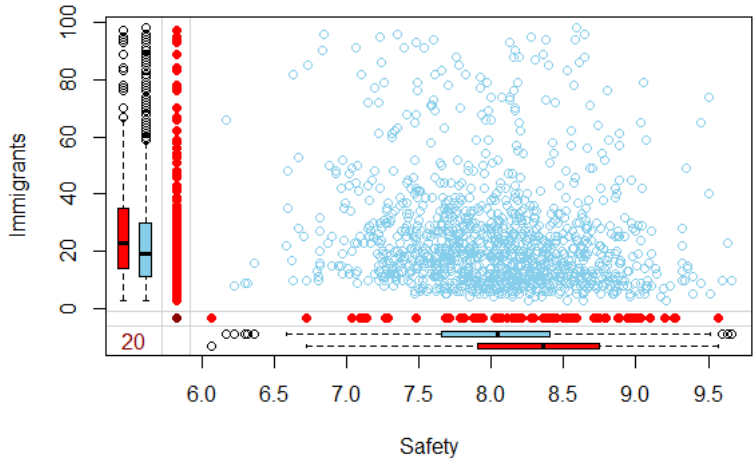
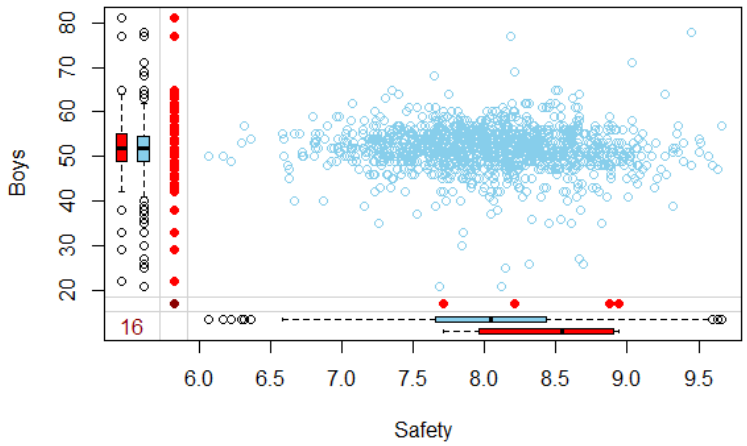


Figure A6: Missing data test for Safety against Boys



After the multiple imputation, I pooled these multiple datasets into one by taking its median value (which keeps variability better than taking mean).

When removing the independent schools there are 1070 public schools left. I then remove the other missing data by listwise deletion as there are fewer than 6.9 % missing for these covariates of *student per teacher* and *certified teachers*. These variables do not seem to have any strong systematic pattern in their missingness when comparing the plots below.

Figure 7A: Missing data test of Student per Teacher against Parents Education

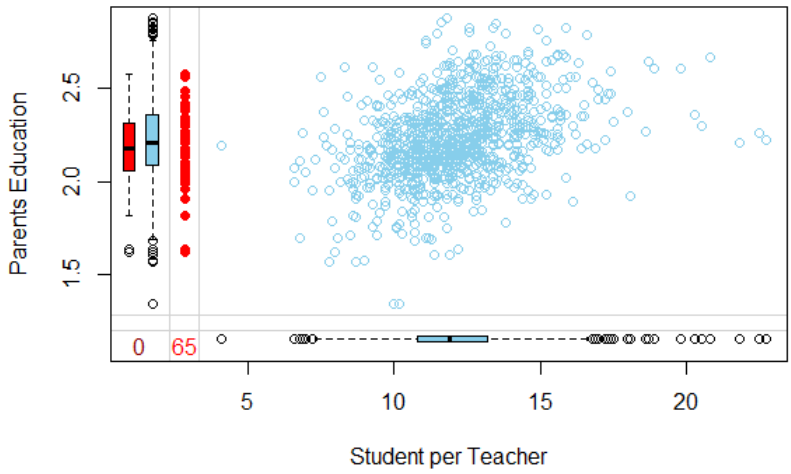


Figure 8A: Missing data test of Students per Teacher against Immigrants

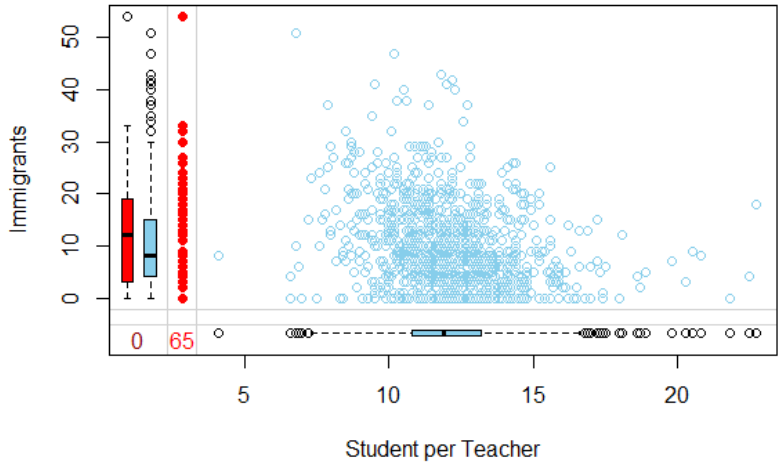


Figure 9A: Missing data test of Students per Teacher against boys

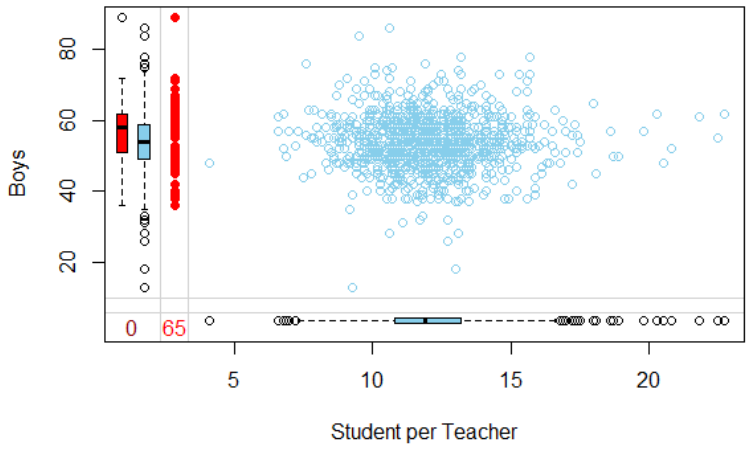


Figure 10A: Missing data test of Certified Teachers against Parents Education

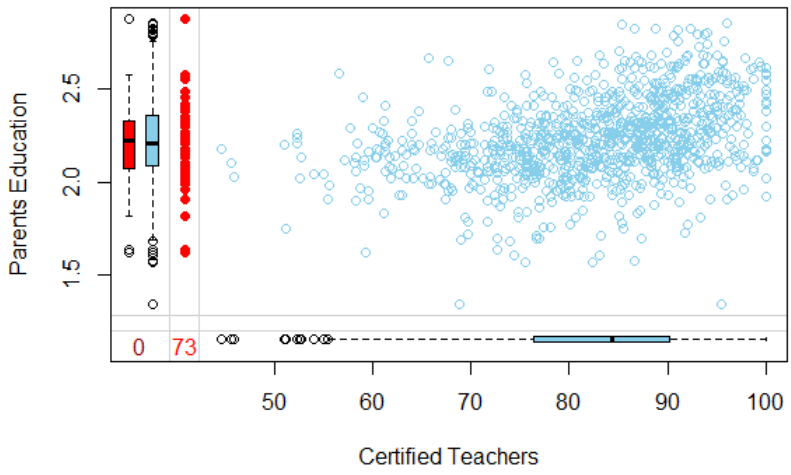


Figure 11A: Missing data test of Certified Teachers against Immigrants

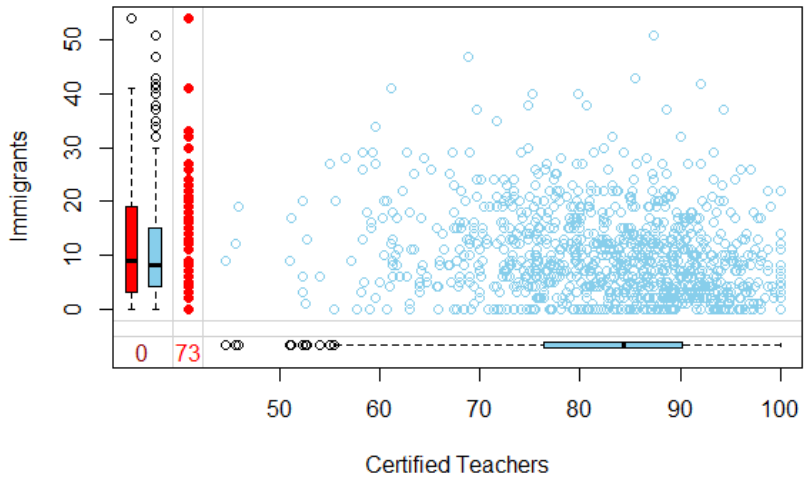
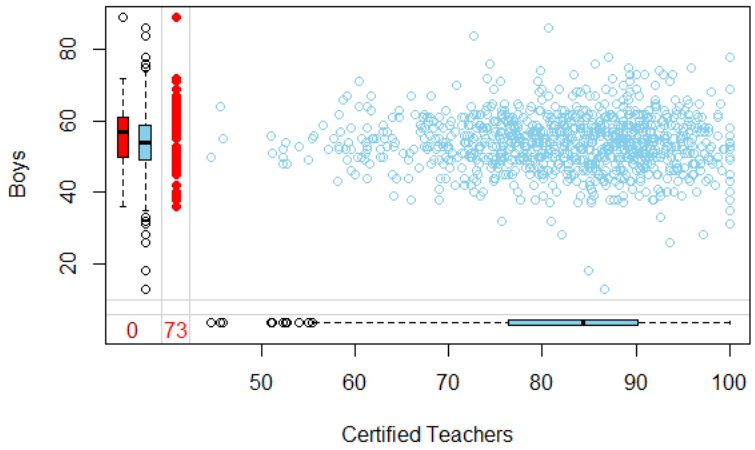


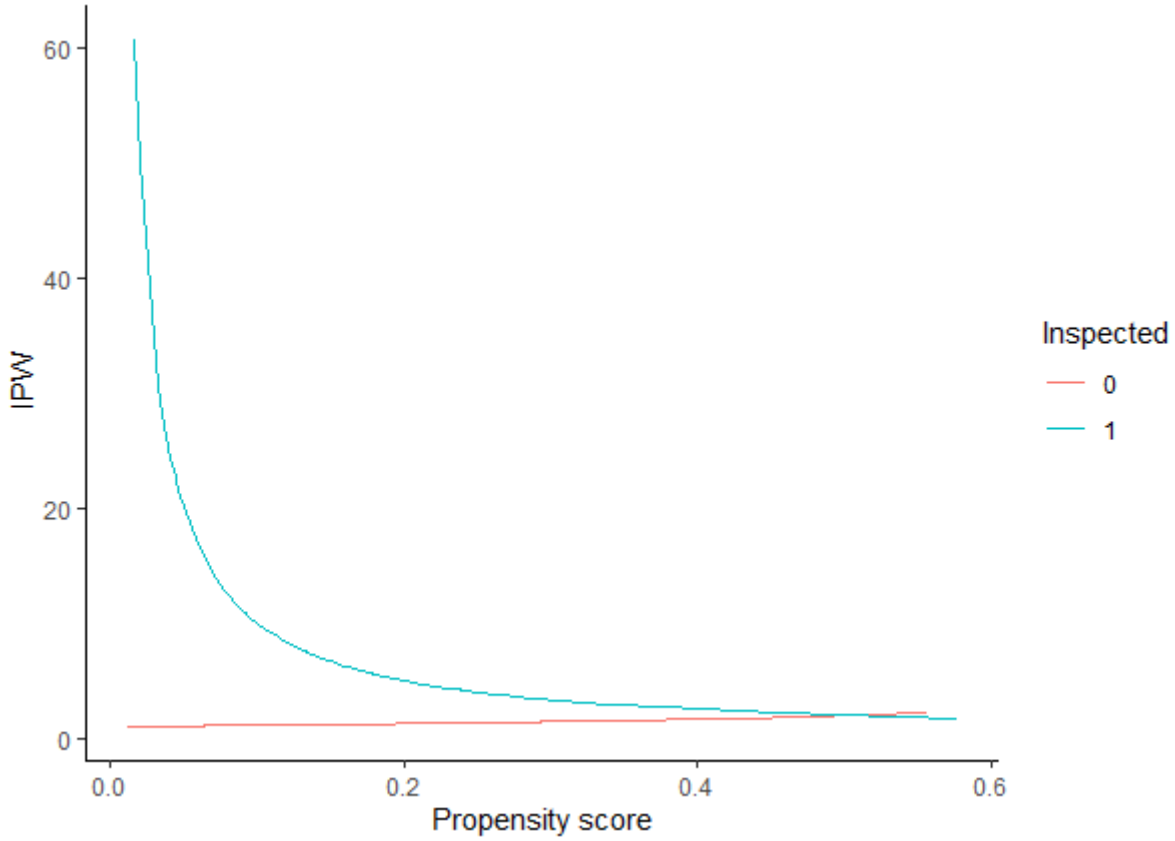
Figure 12A : Missing data test of Certified Teachers against Boys



After the listwise deletion, 976 public schools remained in the sample. Finally, we removed as described in the section "*Strategic and representative sampling*" schools inspected twice (removes 49 observations) and schools' inspection earlier in the cycle 2015 and 2016 (removes 127 observations). After this, we have 800 public schools left.

APPENDIX 2

Figure 13A: Difference in IPW between inspected and non-inspected schools across the propensity score distribution



APPENDIX 3

Test of pre-trends of parallel trends in "empty" models.

Table A1: Empty model of grades to test pre-trends

	Dependent variable:
	Grades
ATT 2014	**-1.047 (3.870)**
ATT 2015	**-5.178 (3.796)**
ATT 2016	**-6.210* (3.695)**
ATT 2017	**-5.665 (3.694)**
ATT 2018	**-4.394 (3.736)**
ATT 2019	**-7.018* (3.790)**
ATT 2020	**-6.936* (3.804)**
ATT 2021	**-4.218 (3.853)**
ATT 2022	**-6.106 (3.858)**
Constant	**213.382*** (0.899)**
Observations	**7,091**
R^2	**0.081**
Adjusted R^2	**0.078**
Residual Std. Error	**21.377 (df = 7071)**
F Statistic	**32.785*** (df = 19; 7071)**
Note:	***p<0.1; **p<0.05; ***p<0.01**

Table A2: Empty model of mathematics test to test pre-trends

	Dependent variable:
	Math
ATT 2014	**0.279**
ATT 2015	**0.002 (0.320)**
ATT 2016	**0.055 (0.312)**
ATT 2017	**-0.456 (0.313)**
ATT 2019	**-0.158 (0.320)**
ATT 2022	**0.012 (0.326)**
Constant	**12.271*** (0.076)**
Observations	**4,959**
R^2	**0.130**
Adjusted R^2	**0.127**
Residual Std. Error	**1.804 (df = 4945)**
F Statistic	**56.652*** (df = 13; 4945)**
Note:	***p<0.1; **p<0.05; ***p<0.01**

Table A3: Empty model of English test to test pre-trends

	Dependent variable:
	Eng
ATT 2014	**0.036 (0.237)**
ATT 2015	**-0.133 (0.232)**
ATT 2016	**-0.174 (0.226)**
ATT 2017	**-0.407* (0.226)**
ATT 2018	**-0.205 (0.237)**
ATT 2019	**-0.149 (0.233)**
ATT 2022	**-0.209 (0.236)**
Constant	**14.786*** (0.055)**
Observations	**5,605**
R^2	**0.043**
Adjusted R^2	**0.041**
Residual Std. Error	**1.306 (df = 5589)**
F Statistic	**16.893*** (df = 15; 5589)**
Note:	***p<0.1; **p<0.05; ***p<0.01**

APPENDIX 4: Bad controls

In this appendix the “Bad controls” test I performed where a similar double-robust difference-in-difference model is performed on the time-varying covariates.

Figure 14A: Effect of inspection on average level of parent’s education over time

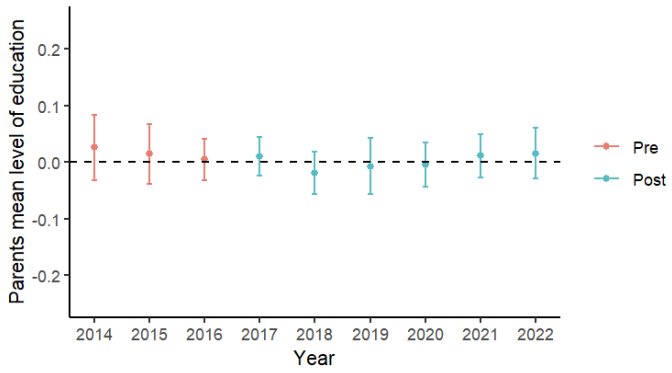


Figure 15A: Effect of inspection on the share of immigrants over time

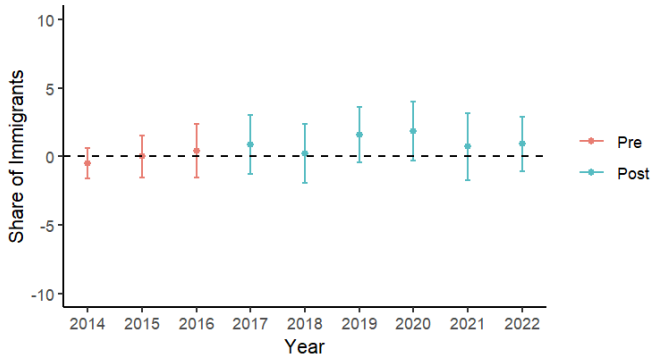
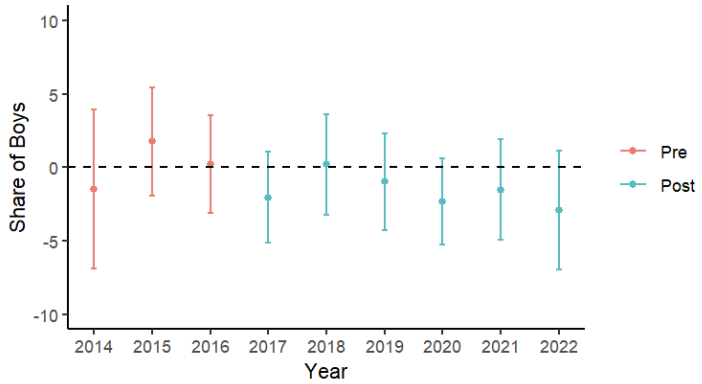


Figure 16A: Effect of inspection on the share of boys over time



APPENDIX 5

In this appendix the main specifications are provided in full. The tables display the models on grades, mathematics test and English test in that order. The columns from left to right display an empty difference-in-difference model, a weighted model to condition on covariates, a weighted model with time-varying controls (double robust).

Table A4: Model results on grades from an empty model, a weighted model and a doubly robust model, from left to right

	Empty	**Weighted**	**Doubly-robust**
Year 2014 ATT (SE)	**-0.047 (0.092)**	**0.056 (0.081)**	**0.051 (0.089)**
Year 2014 CI	[-0.283, 0.189]	[-0.153, 0.265]	[-0.18, 0.283]
Year 2015 ATT (SE)	**-0.185 (0.102)**	**-0.007 (0.082)**	**-0.087 (0.082)**
Year 2015 CI	[-0.448, 0.077]	[-0.219, 0.205]	[-0.299, 0.125]
Year 2016 ATT (SE)	**-0.046 (0.117)**	**0.043 (0.102)**	**0.101 (0.096)**
Year 2016 CI	[-0.346, 0.253]	[-0.221, 0.307]	[-0.148, 0.349]
Year 2017 ATT (SE)	**0.024 (0.143)**	**0.048 (0.134)**	**-0.116 (0.119)**
Year 2017 CI	[-0.343, 0.392]	[-0.3, 0.396]	[-0.424, 0.192]
Year 2018 ATT (SE)	**0.082 (0.101)**	**-0.004 (0.088)**	**-0.05 (0.088)**
Year 2018 CI	[-0.179, 0.342]	[-0.232, 0.225]	[-0.278, 0.178]
Year 2019 ATT (SE)	**-0.036 (0.12)**	**-0.135 (0.09)**	**-0.098 (0.085)**
Year 2019 CI	[-0.344, 0.271]	[-0.368, 0.099]	[-0.318, 0.122]
Year 2020 ATT (SE)	**-0.033 (0.131)**	**-0.03 (0.126)**	**0.09 (0.123)**
Year 2020 CI	[-0.368, 0.303]	[-0.357, 0.297]	[-0.229, 0.409]
Year 2021 ATT (SE)	**0.089 (0.127)**	**0.08 (0.092)**	**0.244* (0.09)**
Year 2021 CI	[-0.236, 0.415]	[-0.158, 0.318]	[0.011, 0.476]
Year 2022 ATT (SE)	**0.005 (0.138)**	**0.024 (0.114)**	**0.225 (0.109)**
Year 2022 CI	[-0.348, 0.358]	[-0.271, 0.319]	[-0.058, 0.508]

P-value pre-trend test	0.04118	0.92163	0.6334
N	800	800	800

Table A5: Model results on mathematics exam from an empty model, a weighted model and a doubly robust model, from left to right

	Empty	**Weighted**	**Doubly-robust**
Year 2014 ATT (SE)	**0.14 (0.133)**	**0.096 (0.129)**	**0.114 (0.138)**
Year 2014 CI	[-0.187, 0.466]	[-0.209, 0.4]	[-0.231, 0.46]
Year 2015 ATT (SE)	**-0.139 (0.223)**	**0.124 (0.166)**	**0.1 (0.151)**
Year 2015 CI	[-0.685, 0.408]	[-0.269, 0.518]	[-0.277, 0.477]
Year 2016 ATT (SE)	**0.027 (0.163)**	**-0.053 (0.146)**	**-0.017 (0.136)**
Year 2016 CI	[-0.373, 0.426]	[-0.398, 0.292]	[-0.357, 0.324]
Year 2017 ATT (SE)	**-0.256 (0.13)**	**-0.132 (0.124)**	**-0.169 (0.11)**
Year 2017 CI	[-0.574, 0.062]	[-0.425, 0.16]	[-0.445, 0.106]
Year 2019 ATT (SE)	**-0.107 (0.131)**	**-0.06 (0.094)**	**-0.002 (0.092)**
Year 2019 CI	[-0.428, 0.214]	[-0.283, 0.162]	[-0.233, 0.23]
Year 2022 ATT (SE)	**-0.022 (0.108)**	**0.089 (0.096)**	**0.226* (0.086)**
Year 2022 CI	[-0.287, 0.243]	[-0.138, 0.316]	[0.011, 0.441]
P-value pre-trend test	0.51609	0.49239	0.4825

N	800	800	800

Table A6: Model results on English exam from an empty model, a weighted model and a doubly robust model, from left to right

	Empty	**Weighted**	**Doubly-robust**
Year 2014 ATT (SE)	**0.036 (0.173)**	**0.001 (0.194)**	**0.019 (0.214)**
Year 2014 CI	[-0.416, 0.488]	[-0.492, 0.494]	[-0.51, 0.547]
Year 2015 ATT (SE)	**-0.169 (0.151)**	**0.029 (0.121)**	**0.004 (0.126)**
Year 2015 CI	[-0.561, 0.224]	[-0.278, 0.337]	[-0.307, 0.316]
Year 2016 ATT (SE)	**-0.041 (0.123)**	**0.112 (0.117)**	**0.181 (0.121)**
Year 2016 CI	[-0.363, 0.28]	[-0.184, 0.408]	[-0.116, 0.478]
Year 2017 ATT (SE)	**-0.232 (0.124)**	**-0.236 (0.154)**	**-0.344 (0.149)**
Year 2017 CI	[-0.556, 0.092]	[-0.626, 0.154]	[-0.711, 0.023]
Year 2018 ATT (SE)	**-0.031 (0.187)**	**-0.068 (0.192)**	**-0.113 (0.182)**
Year 2018 CI	[-0.518, 0.457]	[-0.555, 0.42]	[-0.561, 0.336]
Year 2019 ATT (SE)	**0.026 (0.182)**	**-0.131 (0.159)**	**-0.048 (0.158)**
Year 2019 CI	[-0.45, 0.501]	[-0.534, 0.271]	[-0.437, 0.34]
Year 2022 ATT (SE)	**-0.034 (0.163)**	**-0.157 (0.155)**	**0.11 (0.158)**
Year 2022 CI	[-0.458, 0.39]	[-0.549, 0.235]	[-0.28, 0.5]
P-value pre-trend test	0.36018	0.84826	0.65016
N	800	800	800

APPENDIX 6: Balanced panel test

In this appendix the test when balancing the panel is performed.

Figure 17A: Dynamic effect of being Inspected on Grades on a balanced panel

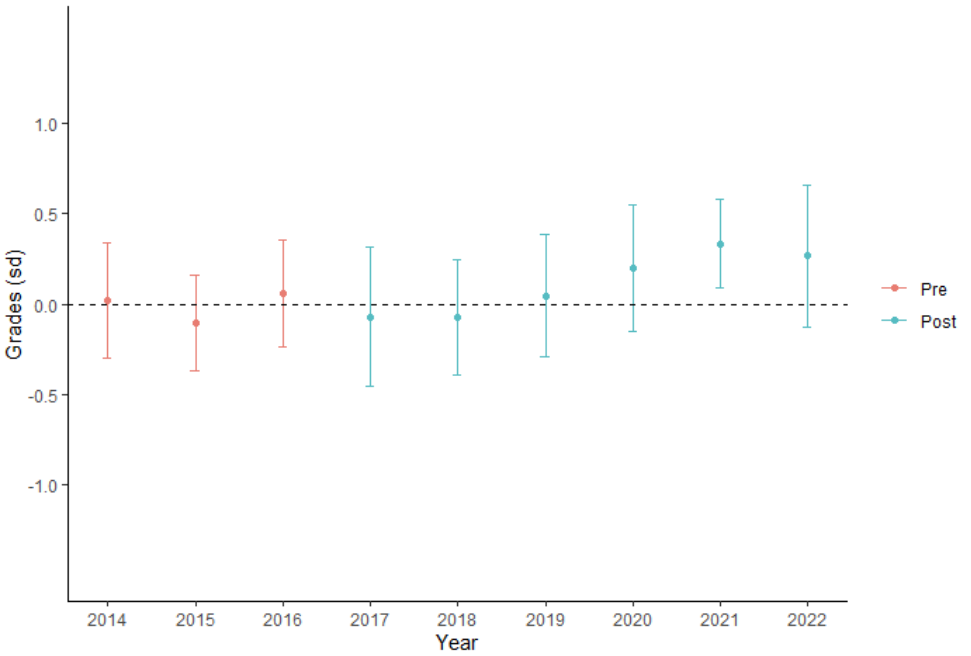


Figure 18A: Dynamic effect of being Inspected on Mathematics test on a balanced panel

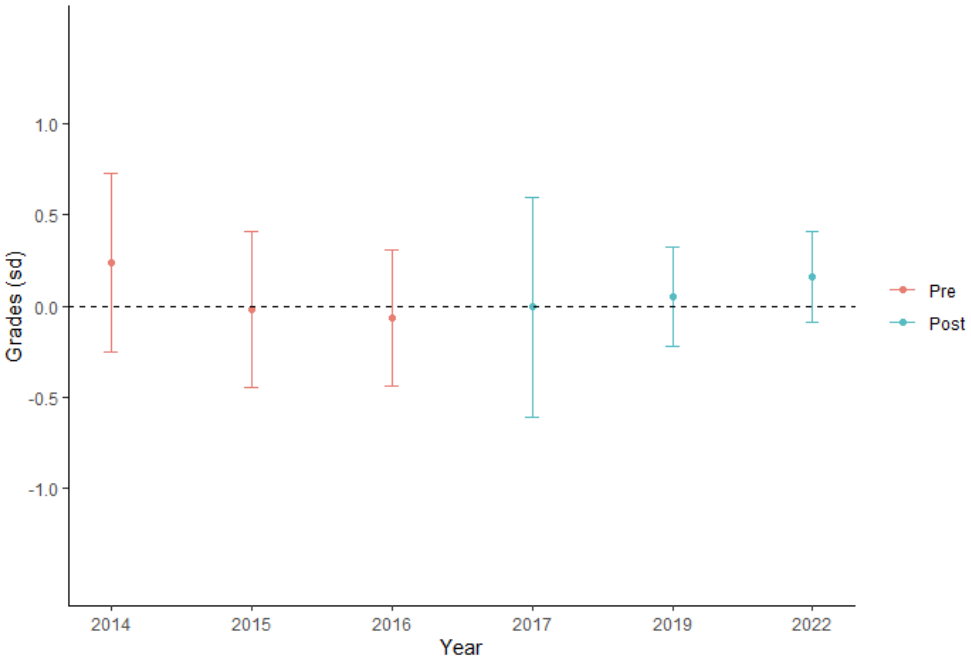


Figure 19A: Dynamic effect of being Inspected on English test on a balanced panel