Beyond Exclusion: The Role of the Causal Effect of Testing on Attendance on the Day of the Test

Magdalena Bennett[‡] **Christopher Neilson**[§] **Nicolás Rojas Souyet**[*]

This version: January 7th, 2026

Abstract

High-stakes testing is widely used across educational systems, shaping accountability, resource allocation, and school choice. While skewed attendance may undermine these goals, little is known about its impact beyond the exclusion of low-performing students. Using an event-study design and rich administrative data from Chile, we examine how testing affects student attendance across grades and performance levels. We find highly heterogeneous effects—ranging from -8 to 4 percentage points across the GPA distribution—with negative impacts only among younger, low-performing students and positive impacts for students above the bottom of the GPA distribution. Using survey data, a separate event study, and exemption records, we rule out student-exemption policies and test anxiety as mechanisms, while communication, prizes, and grading incentives help explain heterogeneous attendance responses. Finally, we discuss the limits of multiple imputation and propose a machine-learning approach to identify schools potentially engaging in gaming behavior, enabling targeted oversight instead of blunt penalties commonly used in the U.S.

[*]Columbia University. [‡]Meta, [§]Yale University. Correspondence to: nicolas.rojas@tc.columbia.edu. We thank seminar and conference participants at the Microsoft Research Seminar (invited talk), the American Causal Inference Conference, the Causal Inference Seminar at the Salem Center, IROM Brown Bag at U.T. Texas, the Methods Talk - Population Research Center (U.T. Texas), the Workshop on Economics of Education at Universidad de los Andes (Valle Nevado), the PRIISM Conference, the AEFP Annual Conferences (Denver; Baltimore), the SREE Annual Conference, and the Ingeniería UC Seminar at the Pontificia Universidad Católica de Chile for helpful comments and suggestions. We also thanks Carolina Flores from the former Research Division at the Quality of Education Agency of Chile at that time for considering our questions on the communication between students, parents, and schools and the Learning Achievement Evaluation Division at the same institution for detailed information about changes in SIMCE regulation. Finally, we also thank the Industrial Relations Section at Princeton University for providing computing resources during the early stages of this project.

1 Introduction

Educational testing has expanded worldwide, reflecting a broader trend toward enhanced accountability for public services across numerous countries (OCDE, 2013; OECD, 2023). U.S. large urban districts offer a notable example, with students taking an average of 112 standardized tests between kindergarten and graduation (Hart et al., 2015). When tests carry high stakes, attendance disparities and schools' incentives to manipulate results interact, obscuring what test scores truly measure. This raises concerns about fairness and accuracy, with implications for school choice, accountability, and the allocation of resources.

Previous research has examined strategies used by schools to exclude lower-performing students. These tactics include inducing parents to withdraw their children from testing (Coelli and Foster, 2024), reclassifying students as having disabilities (Cullen and Reback, 2006; Figlio and Getzler, 2002; Jacob, 2005), applying disciplinary measures to strategically exclude low performers (Figlio, 2006), or increasing student absences in response to accountability policies (Cilliers et al., 2019; Feigenberg et al., 2019; Hofflinger and Hippel, 2020). However, much less attention has been given to non-representative attendance patterns involving both low- and high-achieving students. One notable exception is Cuesta et al. (2020a), which examines the welfare consequences of score differences before and after imputing missing test data using GPA. Still, its analysis of the causal effect of testing on attendance is largely descriptive and limited to a single grade and year. How does high-stakes testing causally affect the skill composition of test-takers, and what policy levers can mitigate these distortions?

In this article, we study the impact of high-stakes testing on the skill composition of test-takers in Chile, where high-stakes accountability has shaped school closures, teacher bonuses, and school choice for more than three decades. Using hundreds of millions of daily attendance records for tested and non-tested students from 2011–2017, we estimate the effect of testing on attendance across academic performance levels and grade levels using an event-study design. We then discuss the limitations of multiple imputation and propose a machine-learning approach to identify schools potentially engaging in gaming behavior, enabling targeted oversight instead of blunt Every Student Succeeds Act–style penalties.

We divide our results into three parts: event-study results, an analysis of potential mechanisms, and a empirical analysis on the limitations of multiple imputations along with the results from our machine-learning predictive approach.

Our event-study results reveal that the effect of high-stakes testing on attendance is highly heterogeneous across grades, with younger students being substantially more affected than older students. Among second graders, the positive impact on higher-performing students (around 3 percentage points, p.p.) is nearly ten percentage points larger than the negative impact on lower-performing students (around −7 p.p.), but this gap narrows to roughly one percentage point by tenth grade. Notably, the positive impact on attendance persists among higher-performing students across grades, while the negative impact among lower-performing students quickly fades out by sixth grade. Thus, student exclusion does not appear to play a meaningful role among students beyond fourth grade, as lower-performing students are no longer negatively affected. Importantly, these results are masked in aggregate estimates: when performance heterogeneity is not considered, the impact of the test day on attendance is generally positive (around 1 p.p.) and consistent with government reports.

We study four mechanisms driving these results. First, we analyze the role of exemption legislation—an important feature of accountability system design studied in prior research (Cullen and Reback, 2006; Figlio and Getzler, 2002; Jacob, 2005)—by estimating event studies with and without exempted students for 2nd and 4th grade. The possibility of formally excluding students with certified permanent special needs does not appear to drive our results: estimates with and without exempted students are nearly identical.

Second, we study whether selective non-participation consistent with a disutility of testing helps explain our results, a behavior related to test opt-out movements in the U.S. (Paladino, 2020), using an event-study design that estimates the impact of a no-stakes version of the test on attendance. In contrast to high-stakes testing, we find precisely estimated null effects across grades, performance levels, and years, indicating that testing *per se* is unlikely to drive changes in the composition of the tested student pool.

Third, we examine the role of communication among students, parents, and schools. Survey evidence indicates that high-performing students are more likely to be aware of test dates, to communicate that information to their parents, and to be better prepared for the test. These patterns are present across grades, suggesting that information plays a key role in increasing attendance among high-performing students.

Fourth, we study the role of prizes and grading incentives. Survey evidence indicates that low-performing students are much more aware of teachers' promises of rewards and "good grades" if the school obtains a high test score[1]. These patterns are present across grades and are particularly pronounced among younger students, with a 34-percentage-point difference between low- and high-performing students—roughly double the gap observed among older students. Together, these findings suggest that prizes and grading incentives could be used—directly or indirectly—to exclude younger low-performing students.

Finally, to address non-representative attendance, we begin by replicating Cuesta et al. (2020a)'s cross-validation procedure to show that multiple imputation methods used in Australia and proposed by Cuesta et al. (2020a) can still be far from true values due to missing-data variability. This leads us to discuss school penalization strategies, which are required in the U.S. to increase attendance (U.S. DOE, 2016; U.S. DOE, 2017). Using a combination of mixed models and gradient boosting (Sigrist, 2022a; Sigrist, 2022b) and daily attendance data leading up to the test day, we forecast attendance on the test day as if it were a typical day. Subsequently, we compare these predictions with actual observed attendance at the school level, enabling us to identify schools potentially engaged in gaming behavior and describe their characteristics. We show that,

[1] Chile do not report individual reports.

in our sample, around 20% of schools appear to engage in strategic behavior and are three times as likely to be at risk of school closure due to accountability policies, despite having a similar share of vulnerable students as non-strategic schools. Across-the-board penalization could end up sanctioning 1/3 of all schools whose attendance falls below 95% because of student composition rather than strategic manipulation, a concern in at least 24 U.S. states where participation below this threshold triggers accountability consequences (Katanyoutanant et al., 2021).

This paper makes several contributions to the literature on school accountability. First, it offers a departure from the predominant focus on the exclusion of low performers from test scores, underscoring the pivotal role of students who are not at the bottom of the performance distribution (Coelli and Foster, 2024; Cullen and Reback, 2006; Cilliers et al., 2019; Feigenberg et al., 2019; Figlio and Getzler, 2002; Figlio, 2006; Hofflinger and Hippel, 2018; Hofflinger and Hippel, 2020; Jacob, 2005). Second, unlike previous studies that focus on a limited set of grades (Cuesta et al., 2020a; Feigenberg et al., 2019; Hofflinger and Hippel, 2018), we examine non-representative patterns from early elementary years through tenth grade, highlighting the potential drawbacks of testing younger students and showing that the negative impact on attendance quickly fades out by sixth grade. This finding is of particular relevance to the ongoing policy debate of the re-implementation of testing in 2nd grade in Chile (Mercurio, 2024). Third, by contrasting high- and low-stakes testing environments, we show that attendance responses are driven by accountability incentives rather than by testing alone. This finding challenges preference-based interpretations of non-participation commonly invoked in discussions of test opt-out behavior in the U.S. and underscores the role of institutional design in shaping participation (Paladino, 2020). Fourth, we provide new evidence on the limitations of imputing test scores to address missing students, showing that even when imputation reduces bias in expectation, it can generate substantial school-level error due to missing-data variability. This extends the debate in Cuesta et al. (2020a) on the use of imputation in accountability systems. Lastly, we introduce a prediction model that aids in identifying potential gaming behaviors and provides a framework for more targeted imputation strategies and participation-based oversight under accountability systems such as ESSA in the U.S. (U.S. DOE, 2016; U.S. DOE, 2017).

The rest of the article is organized as follows. Section 2 describes the policy context and the data. Section 3 details our empirical strategy and Section 4 our data. Section 5 presents our results and discusses potential mechanisms. Section 6 discusses imputations and presents our prediction model strategy. Finally, Section 7 concludes.

2 High-stakes Tests and Attendance in Chile

2.1 Background and General Context

Chile has administered mandatory, high-stakes standardized tests annually for more than 30 years. The SIMCE[2] assessments are administered by the Education Quality Agency, also referred to as the AGCE (*Agencia de Calidad de la Educación*), an institution under the Ministry of Education, in all K–12 schools and cover core areas of the national curriculum, including mathematics, reading, writing, history and geography, and science, in 2nd, 4th, 6th, 8th, and 10th grades. The yearly testing calendar has changed the number of evaluations over time, leading, for example, to the implementation and subsequent suspension of the 2nd-grade reading test.

There are three main reasons why SIMCE tests carry high-stakes consequences. First, results are widely accessible within the context of a nationwide voucher system that started in the early 80's. Second, they are tied to teacher bonuses in publicly funded schools since 1996. Finally, schools may be closed as a result of poor test performance since 2012[3]. Despite their consequences, SIMCE results are always reported at the school level and never at the student or teacher level, even in the case of teacher bonuses.

Along with the regular SIMCE test, the Education Quality Agency also administers no-stakes tests to pilot new questions and conduct psychometric procedures to ensure test comparability over time, among other purposes. The pilot may be administered on the same day as the regular SIMCE, but in some cases the government conducts the test outside the regular schedule, in a different week or even in a different year. This test is taken by a sizable number of schools—typically involving around 10,000 students, and 300 schools in basic education—and is designed to be administered in a manner as similar as possible to the regular SIMCE in terms of providers, procedures, etc (see, for example, Appendix A). The key distinctive feature of this test is that it carries no consequences: results are neither published nor calculated.

2.2 SIMCE Attendance and Exclusion Policies

The Education Quality Agency's main tool to promote attendance is sending a pre-filled letter to help schools communicate and encourage attendance on the day of the test. The letter does not mention exclusion and emphasizes that student attendance is important for the school because some results may not be reported if absenteeism is sufficiently high, a situation that rarely occurs[4] (see Figure A1 in Appendix A for two letter

[2] Originally, SIMCE referred to the System for Measuring Education Quality (*Sistema de Medición de la Calidad de la Educación*). Over time, however, it ceased to function as a system in its original sense, as designed in 1988, and the term is no longer used as an acronym. Today, SIMCE refers solely to the standardized test itself and is officially written as 'Simce' rather than in all capital letters. Throughout this paper, we follow the convention used in the economics and education literature and refer to the test as SIMCE.

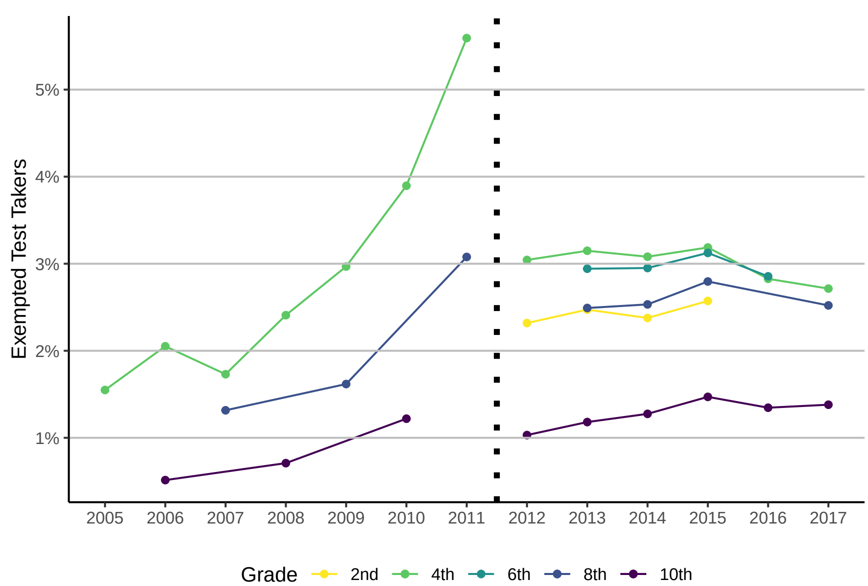
[3] Even though the Preferential Education Subsidy Law, also called the SEP law (or *Subvencion Escolar Preferencial*), for more details see Neilson (2025)) was implemented in 2008, establishing the possibility of school closure, and schools were classified in the category that could lead to closure ("En Recuperacion") for the first time in 2012.

[4] Mean test scores are not reported only if there is one single student who took the test, and performance standards are not reported for fewer than 10 students.

examples). The use of this tool has never been formally evaluated, in part because the Education Quality Agency generally reports that attendance on the day of the test is similar to, or higher than, attendance on a regular school day; however, as we show later in the paper, this aggregate similarity masks substantial changes in attendance composition on the day of the test.

SIMCE tests have a formal exemption policy that legally excludes students with certified disabilities from official school score calculations, while still expecting these students to attend school on the test day[5]. In 2011, educational authorities found that roughly half of the certificates issued did not comply with existing regulations. Beginning in 2012, they implemented a more stringent system to regulate exemptions, in part to align testing practices with the expansion of higher-stakes accountability policies. Justifications for exemptions began to be formally validated. The process requires certificates to be issued by registered health professionals, and the reasons for exemptions were narrowed to permanent special educational needs, such as blindness or autism spectrum disorder. Students should still take a test that can have special accommodations, such as more time to respond, tests with a larger font size to facilitate reading, or tests available in a special digital or physical format. As shown in Figure 1, the number of exemptions declined across all grades, especially in 4th grade, where previously rising exemption rates flattened—suggesting that some degree of manipulation may have occurred before the introduction of certificate validation.

Figure 1: The Rise and Fall of Student Exemptions after the 2012 Regulation



Notes. This figure illustrates the rise and fall of student exemptions that allow schools to legally exclude the scores of students with certified disabilities from official SIMCE results. Following the expansion of accountability policies around 2008, the Ministry of Education introduced a regulation beginning with SIMCE 2012 requiring schools to submit certification from registered health professionals in order to exempt a student's score from official reporting. After 2012, exemption rates declined and became more stable over time.

Even though under the Chilean legislation, altering the results of learning assessments constitutes a serious

[5]The other potential exemption applies to non-Spanish-speaking migrant students, such as recent arrivals from Haiti; however, this category is rarely used at that time and affects a negligible share of students according to SIMCE data.

infraction[6], the Education Quality Agency does not have any specific strategy to address this issue other than the exemptions regulation. Between 2005 and 2007, SIMCE scores were marked with symbols indicating that they were not representative of the school enrollment when absent students had worse GPAs than test-takers (see SIMCE (2006), SIMCE (2007) and SIMCE (2008)). According to informal conversations with test officials at that time, SIMCE stopped highlighting potential manipulation because of principals blaming that low-performing students would not come to school anyway.

3 Data Description

We use four main datasets provided by the Ministry of Education and the Education Quality Agency. First, we build a student-level registry of all school–grade combinations that participated in high-stakes or no-stakes SIMCE tests during 2011–2017. Second, we leverage daily attendance data from the System for General Information of Students (SIGE) to construct a student-day panel for all students attending public or private subsidized schools nationwide around SIMCE test dates. Third, we use national student grade point average (GPA) data called Rendimiento, to classify students based on their academic performance relative to peers within the same school and grade. Fourth, we use academic scores and unpublished survey data from SIMCE 2017 to examine whether students at different GPA levels received information, incentives, or grading incentives related to the test and to complement our prediction analysis.

3.1 Students and Schools taking High-Stakes and No-Stakes SIMCE tests

Table 1 describes the set of tests used to examine the impact of testing on attendance. We built this registry under four restrictions. First, we only consider public and private subsidized schools. Unsubsidized private schools do not have daily attendance records, and they represent around 10% of the student body in total. Second, we do not use high-stakes information for 2012 because there is no daily attendance information after June, and the test takes place during the second semester. Third, we only consider grades and schools with more than ten students at the end of the year to be able to classify students in GPA deciles. Finally, whenever a no-stakes and high-stakes test are taken on the same day, we consider this a high-stakes test[7].

Table 1 shows that, in general, more than 300 schools and thousands of students are required to take the no-stakes test, which is critical for statistical power in our event-study analysis for this subsample.

No-stakes tests are carried out annually to update test items, and to preserve the comparability of results across the years. Schools are randomly allocated within a set of nearly 20 strata. These tests are delivered in the most similar way possible to the regular SIMCE in terms of providers, procedures, and notifications to parents, except that the test has no consequences, and the school results are never published or calculated.

[6] Manipulation of test scores is a "major violation" of the quality assurance system law (article 76 of SACLaw (2011)).

[7] This happend in 10th grade 2015, and 6th grade 2014 and 2015. For these cases, no-stakes test are always taken later in the day in a sample of schools.

Table 1: Participation in SIMCE High-Stakes and No-Stakes Tests (2011–2017)

	High-Stakes			No-Stakes		
Grades	Years Tested	Num. Schools	Num. Students	Years Tested	Num. Schools	Num. Students
2nd	2013, 2014, 2015	5,266	628,073	2011	282	12,915
4th	2011, 2013	5,571	1,241,292	-	-	-
5th	-	-	-	2012	314	13,639
6th	2013-2016	5,331	846,274	2011, 2017	726	35,589
8th	2011, 2013-2015, 2017	5,545	1,078,140	-	-	-
10th	2013-2017	2,622	1,005,937	-	-	-
11th	-	-	-	2012	94	7,850
Total	2011, 2013-2017	24,335	5,436,812	2011,2012,2017	1,416	69,993

Notes. This table shows how many schools and students participated in the regular and no-stakes SIMCE between 2011 and 2017 in our panel, which applies four restrictions: 1) we include only publicly funded schools; 2) we exclude 2012; 3) we exclude schools with fewer than 10 test-takers; and 4) we exclude no-stakes tests that took place on the same day as high-stakes regular tests (10th grade in 2013, and 6th grade in 2014 and 2015).

In total, we use data from 5 grades that are tested on 23 occasions between 2011 and 2017 on the regular high-stakes test and another set of 4 grades that are tested on 5 occasions on the no-stakes test. Although the number of students who take the high-stakes test is nearly 5.5 million students in our sample, substantially higher than the no-stakes test, we have observations from almost 70,000 students in the no-stakes sample.

3.2 Attendance and GPA

In combination with our high-stakes and no-stakes registry, we use the daily SIGE attendance records and the national GPA registry that every school must fill at the end of the year to build student-day panels. Table 2 and Table 3 present descriptive statistics of the combination of both data for high-stakes and no-stakes tests.

The SIGE attendance registry has existed since 2011 and records daily attendance to determine government voucher payments for all public and private subsidized schools. The data have been collected continuously since 2011, except for the second semester of 2012. Each day of the registry contains around 2.5 million records; hence, the complete dataset includes more than 2 billion observations.

We also use the Rendimiento data set to extract the GPA for every student in our sample. Every year, the performance and academic status of all of nearly 3 million students in the country are recorded in this dataset. We classify students as high and low-performing based on their GPA into five groups. D1, D2, D9, and D10 indicate students who are in the first, second, ninth, and tenth deciles of GPA performance within their schools, while D3D8 refers to students in the third through eighth deciles of GPA performance. Our results do not qualitatively change for finer decompositions of D3D8, but sizable increase the computational burden of the estimations.

Table 2: Percent of Days Attended at School around the Test Day (High-Stakes)

GPA	Tested															Non-Tested			
	2nd Grade			4th Grade			6th Grade			8th Grade			10th Grade						
	B	T	A	B	T	A	B	T	A	B	T	A	B	T	A	B	T	A	N Obs
D1	85	79	81	86	85	86	87	87	86	85	84	85	82	80	80	84	85	84	22,661,166
D2	89	87	86	89	90	90	91	91	90	89	90	89	88	87	87	89	90	89	23,516,693
D3D8	91	93	89	92	95	93	93	95	93	92	94	93	91	91	90	92	92	91	129,148,922
D9	93	96	91	94	97	95	94	97	95	94	96	95	93	94	93	93	94	93	20,358,692
D10	94	97	92	95	98	95	95	97	96	95	97	96	95	95	95	94	95	94	18,528,454
ALL	91	92	88	91	94	92	92	94	92	91	93	92	90	90	90	91	92	91	214,213,927

Notes. (1) This table describes our data for high-stakes tests, illustrating how attendance changes across GPA deciles and in the days around the test day. (2) B: attendance five days before the test day; T: attendance on the test day(s); A: attendance five days after the test day, or four days if the test lasts two days. D1, D2, D9, D10: first, second, ninth, and tenth GPA deciles. D3D8: third through eighth GPA deciles.

Table 3: Percent of Days Attended at School around the Test Day (No-Stakes)

GPA	Tested															Non-Tested			
	2nd 2011			5th 2012			6th 2011			6th 2017			11th 2012						
	B	T	A	B	T	A	B	T	A	B	T	A	B	T	A	B	T	A	N Obs
D1	86	86	84	91	92	91	82	81	77	87	88	84	90	90	89	87	88	86	2,718,371
D2	91	92	89	93	94	93	89	88	84	91	92	89	92	92	91	91	92	90	2,808,562
D3D8	93	93	92	95	95	95	92	91	87	93	94	91	94	94	93	93	94	92	15,396,596
D9	94	94	93	96	96	96	93	93	89	95	95	93	95	96	95	95	95	94	2,438,480
D10	95	95	94	96	97	97	94	94	90	96	96	94	96	96	96	95	96	95	2,263,590
ALL	92	93	91	94	95	95	91	90	86	93	93	91	93	94	93	93	93	92	25,625,599

Notes. (1) This table describes our data for no-stakes tests, illustrating how attendance changes across GPA deciles and in the days around the test day. (2) B: attendance five days before the test day; T: attendance on the test day(s); A: attendance five days after the test day, or four days if the test lasts two days. D1, D2, D9, D10: first, second, ninth, and tenth GPA deciles. D3D8: third through eighth GPA deciles.

Table 2 and Table 3, shows the attendance rate for all five performance groups within a 5-day window around the test for tested and non-tested students in high- and no-stakes tests. B represents before, A represents after, and T is the day of the test. Using this window, we are dealing with more than 200 million observations for all high-stakes tests, which alerts about the computational difficulties of running pooled estimates. At the same time, the large number of observations prevents our analysis from being underpowered when assessing pre-trends in our event-study results (Freyaldenhoven et al., 2019).

Importantly, GPA is related to attendance both before, during, and after the test. There is a nearly 4 percentage point attendance gap between students in D2 and D9, and a nearly 10 percentage point difference between students in D1 and D10. For high-stakes tests, while attendance does not change substantially for non-tested grades around the testing days, it decreases for low-performing students and increases for high-performing students. For no-stakes tests, there is little change in attendance around the testing day.

3.3 SIMCE Scores and SIMCE Survey Data

Finally, we combine SIMCE data with our attendance panel for the exemption analysis, and we combine our panel with SIMCE test scores and survey data to study mechanisms and to conduct predictive model analyses.

In the context section, we already presented exemption data from SIMCE in Figure 1. Because the SIMCE test includes individual-level exemption identifiers, we incorporate these indicators into our attendance panel. This allows us to run causal analyses with and without exempted students.

In 2017, the survey regularly administered to all students taking the SIMCE test included five questions about the testing process in 4th, 8th, and 10th grades. At that time, we proposed to the Education Quality Agency two of these questions, which were included in the survey, and asked whether students received a communication sent to their parents regarding the SIMCE day and whether they were informed in advance that the test was approaching. The remaining three questions relate to whether schools prepare students for the SIMCE test and whether they offer prizes or grading incentives when the school obtains "good results." We merge these data into our panel to analyze the correlation between these measures and students' low- and high-performing status.

One limitation of the survey is that it is restricted to students who took the test. However, it is helpful to illustrate the actions schools are taking concerning the test, how they vary among performance levels, and how they could be affecting attendance.

4 Empirical Strategy

4.1 The Impact of Testing on Attendance

We start our empirical analysis by using an event study to identify the effect of testing on the attendance without considering GPA heterogeneity. This approach is closely aligned with how policymakers typically report descriptive statistics about take-up, by comparing attendance on the test day with attendance on a regular day, without accounting for heterogeneous attendance patterns across performance levels (see AGCE (2015a) and AGCE (2015b)). The following equation describes this statistical model:

$$Y_{isgt} = \sum_{t \neq -5} \tau^t D_{isgt}^{tg^*} + \gamma_t + \alpha_i + \varepsilon_{isgt} \tag{1}$$

Y_{isgt} is a binary variable that indicates the attendance of student i, in school s in grade g on the day t. t is centered around the day of the test: $t = 0$. $t = 1$ is also a testing day if a test lasts two days[8]. Y_{isgt} is regressed on a set of indicators $D_{isgt}^{tg^*} = \mathbb{1}\{T_i = t, G_i = g^*\}$, where g^* is the tested grade. Whenever there is more than one grade taking a test within the time window, we keep only one grade g^* in our specification

[8] All tests span two days except for 2nd grade.

to simplify our event study.

Our analysis spans five days before and five days after the test day. Since day ($t = -5$) is our reference period, D goes from 4 days before to 5 days after the test. To illustrate, if a test is administered on a Tuesday, the estimate τ^t captures the change in attendance between the previous Tuesday (day -5) and the test day t, relative to non-tested grades.

Our model also includes time fixed effects, γ_t, with $t = -5$ as the omitted reference period, which capture counterfactual attendance dynamics common to tested and non-tested grades. Additionally, we include an individual fixed effect, α_i, which accounts for unobservable and time-invariant differences between individuals. Finally, ε_{isgt} represents an idiosyncratic error term. For inference, we cluster standard errors at the school level s to account for shocks that may affect all students within a school.

Our main estimand of interest is τ^0, which captures the average treatment effect of the test day on attendance for any student taking the test.

The key identification assumption is that on the absence of the test, the attendance trends of students in treated grades would follow the same trajectory as students in the same GPA groups but in non-tested grades. These standard parallel assumptions can be indirectly tested by testing whether $\tau^t = 0$ when $T_i < 0$.

4.2 The Impact of Testing on Attendance across Academic Performance Levels

To identify the effect of testing on attendance across academic performance levels, our main empirical strategy, we use the same strategy as before but comparing the test between students in tested and non-tested grades within GPA groups. The following equation describes our main statistical model:

$$Y_{isgt} = \sum_{p=1}^{5} \sum_{t \neq -5} \tau^{pt} D_{isgt}^{ptg^*} + \gamma_{pt} + \alpha_i + \varepsilon_{isgt} \tag{2}$$

In this case, Y_{isgt} is regressed on a set of indicators $D_{isgt}^{ptg^*} = \mathbb{1}\{P_i = p, T_i = t, G_i = g^*\}$, which introduce heterogeneity by GPA group, where $p \in \{D1, D2, D3D8, D9, D10\}$ indexes GPA decile groups. In this case, if a test is administered on a Tuesday, the estimate τ^{1t} captures the change in attendance for the bottom GPA decile between the previous Tuesday (day -5) and the test day t, relative to the bottom GPA decile of non-tested grades.

Unlike the previous model, we include GPA-group-by-event-time fixed effects, γ_{pt}, with $t = -5$ at the omitted reference period, which flexibly capture counterfactual attendance dynamics common to tested and non-tested grades within each GPA group.

In this case, our main estimand of interest is τ^{p0}, which captures the effect of the test day on attendance[9] for group p. We are also particularly interested in $\tau^{D1,0} - \tau^{D10,0}$, which measures the gap in attendance responses between top- and bottom-performing students. This difference can be formally tested within the fully interacted version of our regression model.

The key identification strategy is that on the absence of the test, the attendance trends of students in treated grades would follow the same trajectory of students in the same GPA groups but in non-tested grades. These standard parallel assumptions can be indirectly tested by testing whether $\tau^{pt} = 0$ when $T_i < 0$ for all p.

4.3 The Impact of No-Stakes Testing on Attendance across Academic Performance Levels

Finally, in our mechanisms part, we applied a similar strategy to study the effect of no-stakes testing across academic performance levels. The events-study is the same, with the difference that we include an additional layer of heterogeneity, which is whether the student attend a school that takes the pilot, defined as $v \in \{0, 1\}$. In this case, τ^{ptv} captures the change in attendance for GPA group p between the previous weekday (day -5) and the test day t, relative to the bottom GPA decile in non-tested grades, for schools with pilot status indicated by v.

$$Y_{isgt} = \sum_{v=0}^{1} \sum_{p=1}^{5} \sum_{t \neq -5} \tau^{ptv} D_{isgt}^{ptg^{*}v} + \gamma_{pt} + \alpha_i + \varepsilon_{isgt} \tag{3}$$

Our main estimand of interest is τ^{P01}, which captures the effect of the test day $(t = 0)$ on attendance for group p in schools that take the no-stakes test, defined as $v = 1$.

5 Results

This section is divided into two parts: first, we present our main results on the effects of high-stakes testing on attendance; second, we examine four possible underlying mechanisms: exemptions, test-anxiety, student–parent–school communication, and prizes and grading incentives.

5.1 Results of Testing on Attendance

We find three main results on the effects of high-stakes testing on attendance. First, the impact of the test day on attendance, when performance heterogeneity is not considered, is generally positive and consistent

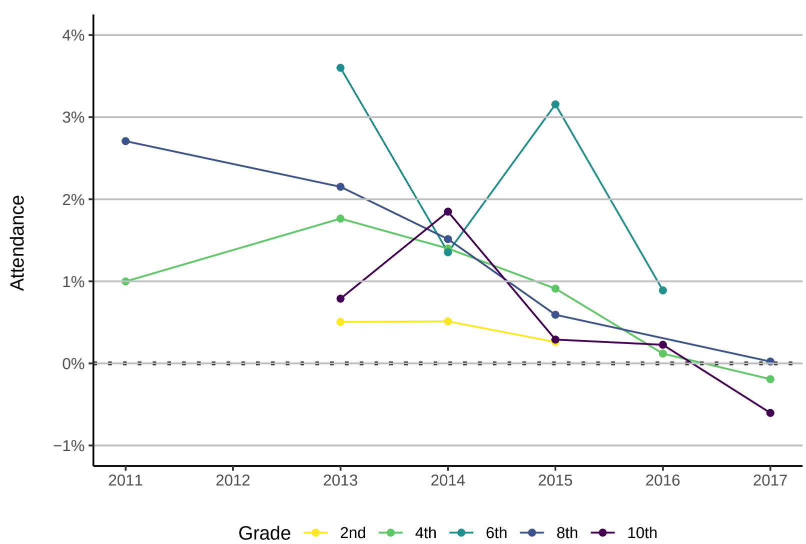
[9] We also show τ^{p1} when the test last two days in our event study plots.

with government reports. Second, these positive aggregate effects mask substantial heterogeneity across grades and performance levels, with sizable negative effects for younger low-performing students but not for older low-performing students. Finally, we find unexpected evidence of negative attendance effects in the days following the test.

Figure 2 shows a positive impact of the test day on attendance for nearly all students taking SIMCE, using an event-study specification that does not account for performance heterogeneity, except a small negative effect for 10th grade in 2017. (see Table 4 and Table A1 in the Appendix for inference). This is consistent with national reports from the Education Quality Agency, which generally present SIMCE attendance relative to a regular school day as a success (see AGCE (2015a) and AGCE (2015b)). For most grades and years in our sample, SIMCE increases attendance by approximately one percentage point, although these positive effects appear to decline over time between 2011 and 2017.

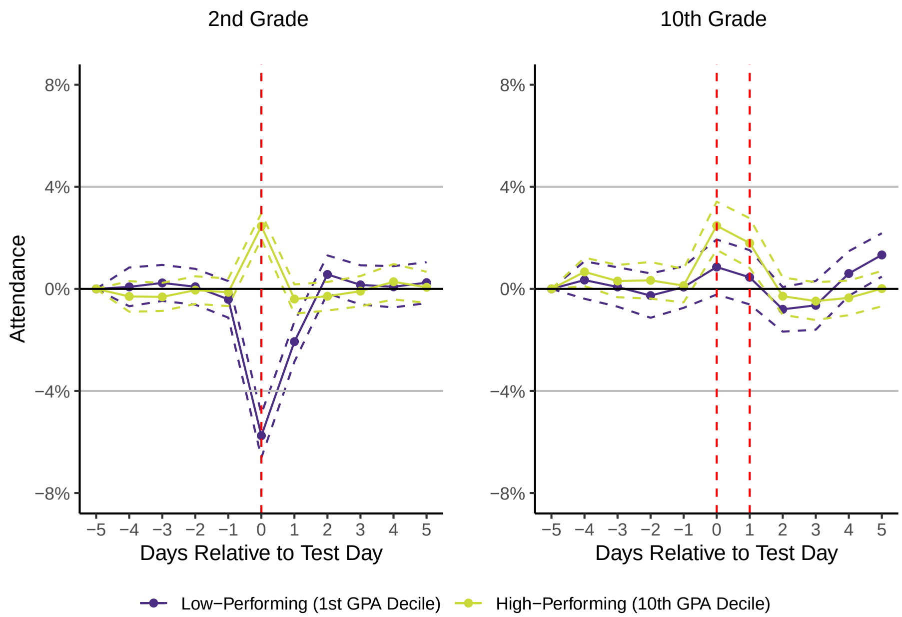
However, Figure 3, Figure 4, and Table 3 reveal that this positive effect masks sizable heterogeneity across grades and performance levels. In Figure 3, we observe an 8 percentage point gap in the impact of testing on attendance between top- and bottom-performing second-grade students in 2014, despite a positive average effect on attendance for all students. This gap is four times smaller for 10th graders in the same year, but remains statistically significant, consistent with the results reported in Table 3 . In addition, Figure 3 shows no evidence of pre-trends for either 2nd or 10th grade in 2014, supporting the parallel trends assumption required for causal inference (see Figure A3 for any other grades in 2014).

Figure 2: Impact of Testing on Attendance



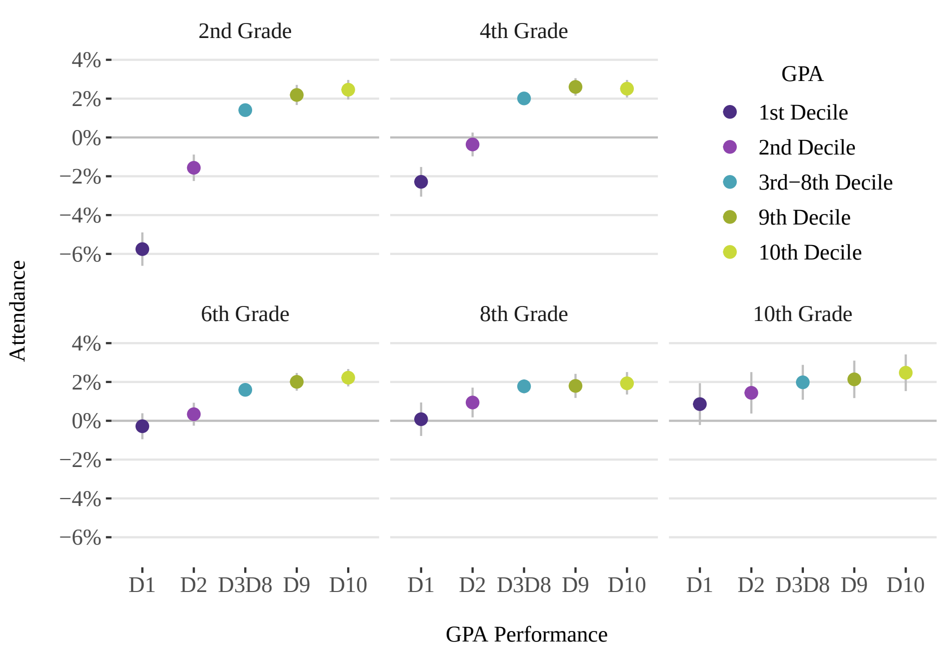
Notes. This figure shows that event-study results that do not account for GPA heterogeneity are predominantly positive: 21 estimates are positive, while only 2 exhibit small negative effects. This illustrates how the heterogeneous effects documented in the figures below are masked when take-up is studied at an aggregate level, which is consistent with how SIMCE attendance results are typically presented. Each dot corresponds to an estimate of τ^0 (see Equation 4.1) from a separate event study; inference is presented in the ALL columns in Table 4 and Table A1 in the Appendix.

Figure 3: Event Study Estimates Around the Test Day (2014)
Low- and High-Performing Students



Notes. (1) This figure illustrates the substantial gap in the impact of testing on attendance between low- and high-performing students among younger grades, relative to older grades, in 2014, and shows no evidence of pre-trends, supporting the parallel trends assumption. It also shows that the attendance gap between grades taking and not taking the test returns to pre-trend levels shortly after the test day. (2) Dashed lines are 95% confidence intervals, and the vertical red lines indicate the test day. In 10th grade, the test was administered over two consecutive days. (3) Pre-trends for all grades in 2014 are in the Appendix.

Figure 4: Impact of the Testing on Attendance across Academic Performance Levels (2014)



Notes. This figure illustrates our main results for 2014, showing substantial heterogeneity in the impact of testing on student attendance. The results indicate that the exclusion of low-performing students accounts for only part of the overall impact. While low-performing students experience negative effects only in younger grades, students who are not at the bottom of the GPA distribution consistently exhibit positive attendance responses across all grades. Vertical lines represent 95% confidence intervals.

Table 4: Impact of Testing on Attendance Across Academic Performance Levels (2013-2015)

Year/Grade	D1	D2	D3D8	D9	D10	D10-D1	ALL
2013							
2nd	-0.06***	-0.02***	0.01***	0.02***	0.03***	0.1***	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
4th	-0.03***	0.00	0.02***	0.03***	0.03***	0.07***	0.02***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
6th	0.02***	0.03***	0.04***	0.04***	0.04***	0.02***	0.04***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
8th	0.00	0.01***	0.03***	0.03***	0.03***	0.03***	0.02***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
10th	0.01*	0.00	0.01***	0.01***	0.01***	0.00	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
2014							
2nd	-0.06***	-0.02***	0.01***	0.02***	0.02***	0.08***	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
4th	-0.02***	0.00	0.02***	0.03***	0.03***	0.05***	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
6th	0.00	0.00	0.02***	0.02***	0.02***	0.03***	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
8th	0.00	0.01**	0.02***	0.02***	0.02***	0.02***	0.02***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
10th	0.01	0.01***	0.02***	0.02***	0.02***	0.02***	0.02***
	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
2015							
2nd	-0.08***	-0.02***	0.01***	0.03***	0.03***	0.11***	0.00*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
4th	-0.02***	-0.01***	0.01***	0.02***	0.02***	0.05***	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
6th	0.02***	0.03***	0.03***	0.04***	0.04***	0.02***	0.03***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
8th	-0.01***	0.00	0.01***	0.01***	0.02***	0.03***	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
10th	0.00	0.00	0.00	0.01	0.00	0.01	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)

Notes. (1) This table shows the main results from Equation 2, except for the last column, which present results from the event-study specification that does not account for heterogeneity, as shown in Equation 4.1. (2) D1, D2, D9, and D10 correspond to the first, second, ninth, and tenth GPA deciles, respectively. D3D8 corresponds to the third through eighth GPA deciles. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the school level. The results for the rest of the results are presented in Table A1 in the Appendix.

Figure 4 compares the impact across all GPA groups and grades in 2014 and shows that the negative impact on attendance fades quickly by sixth grade. Although the test day continues to affect the composition of test-takers, with a gap of close to 2 percentage points between higher- and lower-performing students, this gap is substantially smaller in later grades. Overall, while the test day still generates non-representative attendance, it is much less likely that student exclusion plays a meaningful role among older students, as lower-performing students are no longer negatively affected. In fact, the second decile in 10th grade exhibits a positive and statistically significant increase in attendance on the test day.

Finally, we find unexpected evidence of a negative attendance effect in the days immediately following the test, as shown in Figure 3 and Figure A3 in the Appendix. This effect is larger for low-performing students. Anecdotal evidence suggests that many schools take a break after the SIMCE test. This finding has important implications, as lower-performing students are less likely to return the post-test surveys used to collect non-academic information, as shown in the mechanisms subsection.

5.2 Mechanisms

In this section, we draw on three sources of evidence to discuss four potential explanations. First, using exemption identifiers, we show that although student exemptions have been identified as an important exclusion mechanism in prior research, our event-study results are unchanged when exempted students are dropped from our data. Second, we rule out selective non-participation driven by a disutility of testing as an explanation for our findings. Third, we show that communication among students, parents, and schools may play a key role in increasing attendance for students who are not at the bottom of GPA distribution. Fourth, we present suggestive evidence that prizes and grading incentives may be used—directly or indirectly—to exclude younger low-performing students.

5.2.1 Exemptions

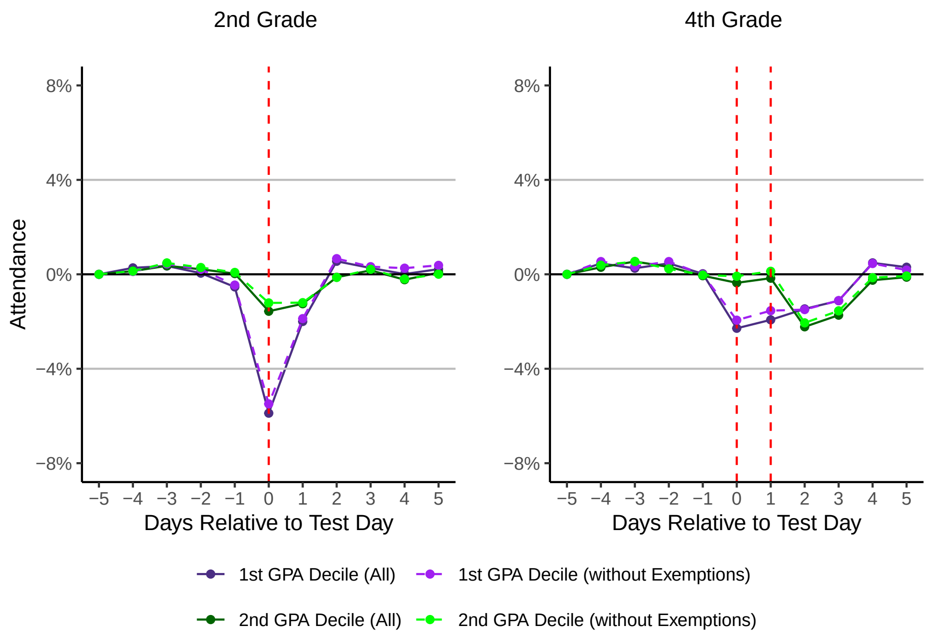
The negative impact of the test day for younger students, as well as the smaller positive impact for older students, may be driven by legislation that determines which students are counted when evaluating student learning. There is substantial evidence that schools respond to such rules by manipulating which students are present or counted on the test day, particularly among lower-performing students (Cullen and Reback, 2006; Figlio and Getzler, 2002; Jacob, 2005). Although exemption certificates may influence reported test outcomes—given that exempted students are predominantly low-performing—our results indicate that this channel does not drive changes in attendance.

As shown in Figure 1 and discussed in Section 2, around 3% of students in grades 2 through 8 have exemption certificates after 2012, and the growth in the share of exempted students has been a concern. Although exempted students are expected to be present on the test day, those with a valid certificate have their test scores excluded from reported results. Because certificates must be presented in advance, students who already know that their scores will not count toward school results may have limited incentives to attend school on the test day, even though the Education Quality Agency prepares special test forms for different types of disabilities. Moreover, exemptions are largely restricted to students with permanent special needs, whose GPA is substantially below that of other students, as shown in Figure A4 in the Appendix. As a result, the exclusion of special-needs students from test score calculations could plausibly contribute to the negative attendance effects observed among low-performing students, particularly at the bottom of the GPA distribution.

To assess this mechanism directly, we follow exempted students over time using SIMCE registries. Figure

5 presents results from an event-study specification estimated with and without exempted students for the first and second GPA deciles in second and fourth grade in 2014. These grades exhibit the largest negative attendance effects and are therefore the most likely to be affected by exemptions. In contrast to previous research, we find no evidence that excluding students from test score calculations affects attendance on the test day.

Figure 5: Event Study with and without Exemptions (2014)



Notes. This figure illustrates that the event-study estimates around the test day do not change when excluding the attendance of students who are exempted. The figure includes only 2nd and 4th grades, because these are the only grades where we find negative impacts on low-performing students.

5.2.2 Test-Anxiety

Another student factor that could explain our results is the disutility of taking a test. There is a body of evidence showing that test anxiety is associated with lower academic performance at every educational level (Seipp (1991), Chapell et al. (2005), Hembree (1988), OECD (2017)). Therefore, low-performing students may suffer more from test anxiety, a form of disutility conceptually related to non-participation under test opt-out movements in the U.S. (Paladino, 2020). We find no evidence that this mechanism plays a role.

To study this mechanism, we leverage the fact that the Education Quality Agency administers a no-stakes test to update test items and preserve the comparability of results across years, as described in Sub-Section 3.1. We then estimate an event-study design for no-stakes tests presented in Sub-Section 4.3.

Table 5 shows our results. Testing alone does not appear to affect attendance on the test day, nor the composition of students present. Although there are a few small positive and negative effects, these appear to be mostly associated with multiple testing. In addition, the aggregated event-study shows some small positive effects on attendance, which are consistent with our previous results. Importantly, most of the null results are estimated with sufficient precision to rule out effects of the magnitude observed for high-stakes tests in most cases.

Table 5: No-Stakes Results

Grade-Year	D1	D2	D3D8	D9	D10	D10-D1	ALL
2nd 2011	-0.01	0.01	0.01*	0.02	0.00	0.01	0.01***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)
5th 2012	0.00	-0.01	0.00	0.01	0.01	0.01	0.00***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)
6th 2011	0.02*	0.01	0.01**	0.01	0.00	-0.02	0.01***
	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)
6th 2017	0.00	0.03	0.01	0.01	0.00	0.00	0.01
	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)
11th 2012	0.00	0.00	0.00	-0.02**	0.00	0.00	0.00
	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)

Notes. (1) This table shows that no-stakes tests have little to no effect on attendance. These estimates come from Equation 4.3, except for the last columns, which present results from the event-study specification that does not account for heterogeneity. (2) D1, D2, D9, and D10 correspond to the first, second, ninth, and tenth GPA deciles, respectively. D3D8 corresponds to the third through eighth GPA deciles. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the school level.

5.2.3 Student–Parent–School Communication

Heterogeneity in communication between schools, parents, and students across GPA levels may help explain why the test does not increase attendance among lower-performing students across grades. Consistent with this mechanism, we find that student–parent–teacher communication is stronger above the bottom of the GPA distribution, particularly among younger cohorts and higher-performing students, suggesting that lower-performing students receive less information or encouragement to attend on test days.

Using data from the 2017 SIMCE student survey, we estimate linear probability models in which the dependent variable is an indicator for whether (i) teachers informed students about the SIMCE test, (ii) the school notified parents about the SIMCE test day, (iii) students attended preparation sessions for the SIMCE test, or (iv) students returned the parent survey used to measure non-academic indicators such as school climate, which accounts for 40% of the non-academic school-climate accountability indicator. The main independent variables are indicators for GPA groups.

Table 6 shows that student-parent-teacher communication is weaker for low-performing students and stronger for high-performing students. The difference is particularly larger for younger students and amounts to a nearly 15 p.p. gap for both questions between the first and last decile of the GPA distribution compared to 3 p.p for older students. The lowest student-parent-school communication is also present for the return of the parent survey that is given to students on the day of the test, which has consequences on the measurement of non-academic indicators used for accountability purposes. In the sixth column of Table 6 we show that low-performing students were around 12 percentage points less like to return parents survey for all grades, compared to high-performing students.

Table 6: Survey Evidence (2017)

	Told	Not	Prep	Prizes	Grades	PSurv
4th Grade						
D1	-0.06***	-0.11***	-0.08***	0.03***	0.14***	-0.07***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
D2	-0.04***	-0.07***	-0.05***	0.03***	0.11***	-0.04***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
D9	0.05***	0.04***	0.04***	-0.04***	-0.14***	0.03***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
D10	0.06***	0.05***	0.05***	-0.05***	-0.2***	0.04***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.89***	0.87***	0.89***	0.55***	0.39***	0.91***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
D10-D1	0.12***	0.16***	0.13***	-0.08***	-0.34***	0.11***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)
8th Grade						
D1	-0.02***	-0.02***	-0.02***	0.04***	0.07***	-0.07***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
D2	-0.01***	-0.01***	-0.01***	0.03***	0.05***	-0.04***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
D9	0.01***	0.00	0.01***	-0.04***	-0.07***	0.03***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
D10	0.01***	0.01***	0.02***	-0.06***	-0.09***	0.05***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.96***	0.86***	0.86***	0.42***	0.26***	0.89***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
D10-D1	0.03***	0.03***	0.04***	-0.1***	-0.16***	0.12***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
10th Grade						
D1	-0.02***	-0.01***	-0.02***	0.03***	0.05***	-0.09***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
D2	-0.01***	0.00	-0.01***	0.02***	0.03***	-0.05***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
D9	0.01***	0.00	0.00	-0.03***	-0.05***	0.04***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
D10	0.01***	0.00	0.00	-0.03***	-0.06***	0.06***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.95***	0.78***	0.82***	0.47***	0.33***	0.84***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
D10-D1	0.03***	0.01***	0.02***	-0.06***	-0.12***	0.15***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)

Notes. Told: Teachers told us about the test. Not: Teachers sent us a parent notification about the SIMCE day. Prep: SIMCE test preparation. Grades: Promized good grades for SIMCE results Prizes: Teachers promised prizes for good results. PSurv: Returned the Parent Survey (among test attendees). All regressions include school fixed effects.

Finally, also in Table 6, we show that low-performing students are less like to have attended a preparation session for the SIMCE test, which something almost all school does in Chile. The difference between top and bottom students is particularly high for younger students with differences that amount almost 13 p.p compared to 4 and 2 p.p for eight and tenth grade. This is also consistent with much less attendance on the days prior to the day of the test as described in Table 2 and Table 3 in the data section.

5.2.4 Prizes and Grading Incentives

It is common practice for Chilean schools to promise grades or grading incentives when the school achieves high SIMCE results, which could, directly or indirectly, induce low-performing students to be absent and high-performing students to be present on the test day. We find that prizes and grading incentives do not appear to be associated with an increase in higher-performing students' attendance, but they are consistent with negative impacts on younger, lower-performing students.

We estimate the same linear probability models as before. This time, the dependent variable is an indicator for whether (i) teachers promised prizes or (ii) grades, if the school obtains a high test score.

Table 6 shows that high-performing students are between 5 p.p. and 20 p.p. less likely to remember having received any of these offers compared to students in the center of the distribution, which suggests that these offers may not be a main driver of the positive impact on attendance for this group. However, younger low-performing students are substantially more likely (14 p.p.) to remember having received grade offers compared to students in the center of the distribution, which is consistent with exclusion patterns. The 34 percentage point gap between top- and bottom-performing students in 4th grade is roughly double the gap observed for students in 8th and 10th grades, which still persists. Overall, the fact that lower-performing younger students are less prepared for the test, have received less communication about it, but remember much more clearly that students will receive a prize or "good grades" suggests that these tools could have been used to directly promote exclusion.

6 Addressing Missing Data: Imputations and Prediction Model Strategy

6.1 Challenges in Multiple Imputation

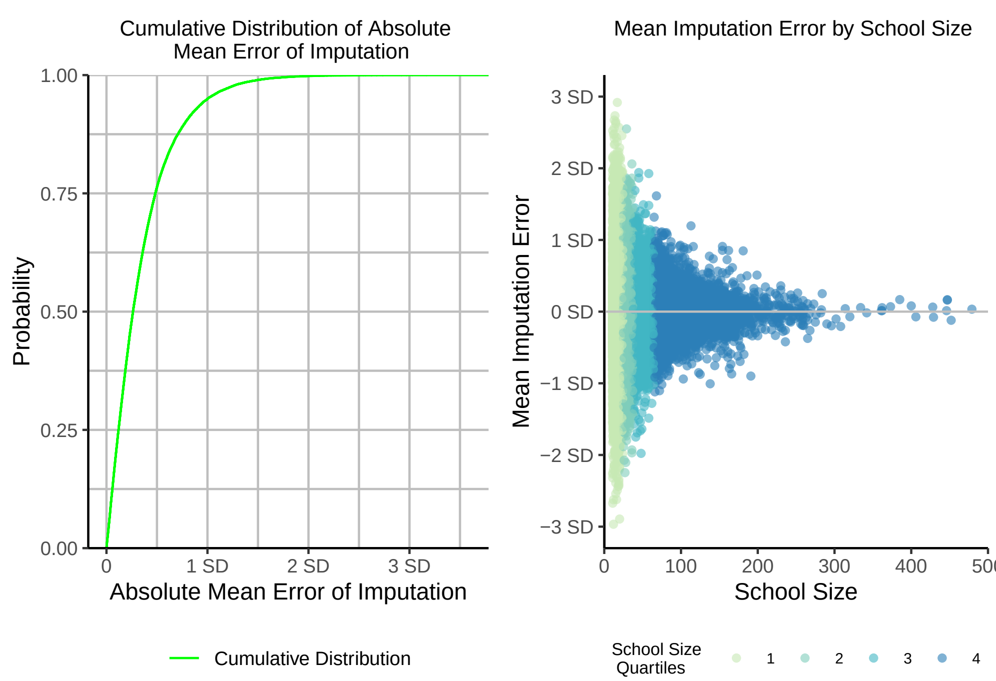
In Table 2 we showed that attendance patterns during the test are not random at all, potentially leading to severe bias problems when ignored (Van Buuren, 2018). Cuesta et al. (2020a) demonstrates that it is indeed the case for Chile, and there are substantial welfare consequences of overlooking absences of student test-scores. Cuesta et al. (2020a) suggests using a multiple imputation procedure for the scores of absent students, as in the NAPLAN test in Australia. However, an unaddressed potential problem with imputation is that even if the imputation model yields unbiased estimates of true score values, school-level scores can still deviate substantially from their true values due to variability induced by missing data. After all, missing scores remain missing.

To illustrate the relevance of this problem, we replicate the cross-validation procedure used in Cuesta et al. (2020b) to evaluate their imputation model and then we calculate the mean absolute error of imputation. The validation procedure consists in working with the complete data set, generating 10% of random absences for each school, a number similar to actual absences, running the multiple imputation model in this new data, and comparing the true values to the imputed values. As in Cuesta et al. (2020b) , the imputed values are, on average close to the true values (see our replication of see Figure A5 in the Appendix).

The problem is that a single school could have 40 students, and if 4 students are missing, the mean prediction for these students could be far away from the their true scores because of the variability of the distribution of plausible values. To examine the magnitude of this problem, we compute the imputation error for each individual student as the difference between the true value and the mean score of all the plausible values and we average the imputation error at the school level.

Figure 5 shows our results. On the left, we show the cumulative distribution of the mean absolute imputation error. For 25% of schools, there is at least a 0.5 standard deviation difference[10] between the imputed score and the true score at the school level. On the right-hand side, we observe that this problem is clearly related to school size: larger schools exhibit much lower imputation error than smaller schools.

Figure 6: School Imputation Error



Notes. (1) These figures show that the difference between observed and predicted scores (i.e., the imputation error), drawn from our replication of the cross-validation exercise in Cuesta et al. (2020b) for 4th-grade mathematics scores in 2014, exhibits substantial variability. (2) We exclude very small schools ($n < 10$).

[10]One standard deviation is defined as the standard deviation of the individual score distribution, which is around 50 SIMCE points.

6.2 A Prediction Model Strategy

Following Van Buuren (2018), preventing missing data may be preferable to imputation. One approach to addressing non-representative participation patterns has involved penalization strategies. In the early 2000s, SIMCE flagged school scores when absent students had lower GPAs than test-takers, but later discontinued the practice, as principals argued that low-performing students would not attend test days regardless[11]. Similarly, in the United States, states are required to address non-representative participation in testing by incorporating participation rates into accountability systems, and many states include penalization strategies when participation falls below 95% for any student subgroup (U.S. DOE, 2016; U.S. DOE, 2017; Katanyoutanant et al., 2021). These experiences raise the question of how non-representative participation should be identified and addressed in practice.

In this subsection, we leverage a data-driven approach to predict attendance on the day of the test as if it were a regular day. This predicted distribution allows us to make a fair comparison of how attendance should look for a given school, given its performance distribution and, thus, identify schools that behave differently from what is predicted. We show that, around 20% of schools appear to engage in strategic behavior and are three times as likely to be at risk of school closure due to accountability policies, despite having a similar share of vulnerable students as non-strategic schools. Across-the-board penalization could end up sanctioning 1/3 of schools whose attendance falls below 95% because of student composition rather than strategic manipulation.

For predicting attendance for each decile at the school-grade level, we leverage the GPBoost algorithm, which combines gradient boosting and Gaussian processes (Sigrist, 2022a; Sigrist, 2022b). This combination allows us to get accurate predictions while taking into account the panel structure of the data. By using all attendance records up to the day before the test, we build a clustered model with random effects at the student level and date level. The model also includes fixed effects by day of the week, school, and grade, as well as siblings' attendance, to capture potential trends or other patters that could be good predictors of absences. Given the size of the dataset, as a first step we restrict the analysis to a single year (2017) and a single testing cohort—grades 1 through 5, of which only 4th grade is tested—in the Metropolitan Region.

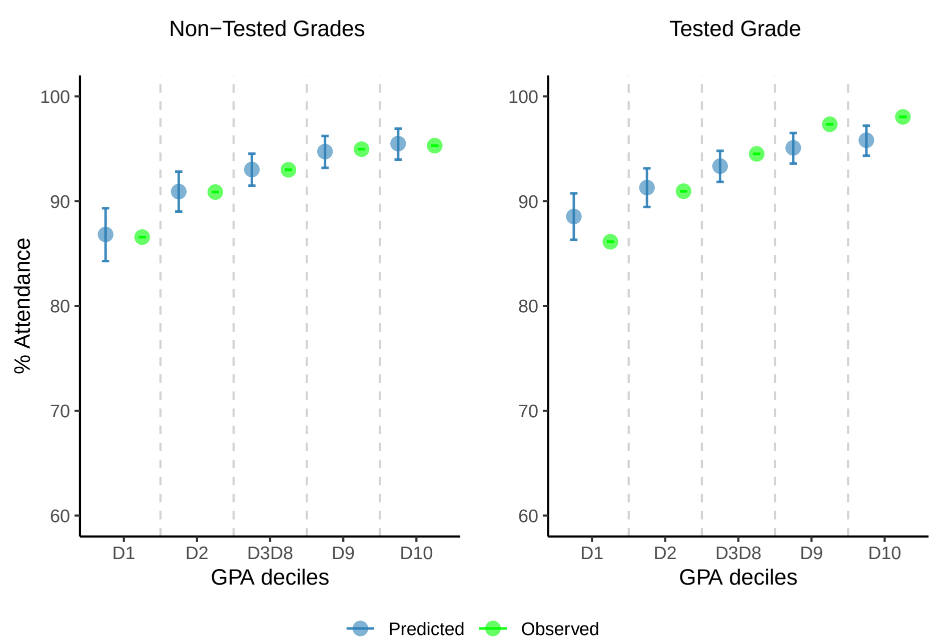
Figure 7 shows our overall predictions for non-tested grades and the tested grade (4th grade). Consistent with our event-study results, the model accurately predicts attendance on a regular day for non-tested grades, but there is a significant difference when we compare our predicted values and observed values for the tested grade.

To identify schools that are potentially engaging in gaming behavior and characterize these types of schools, we use a K-means approach to identify the optimal number of clusters in our data. The variables that we use for this model are the difference between the observed attendance for each decile and the predicted

[11] See Sub-Section 2.2 for more details.

attendance, so each group will have a different behavior with respect to observed and predicted values. Under this approach, the optimal number of clusters we find is 2, and the differences between their observed and predicted values for each decile are depicted on the first four rows of Table 7. We label cluster 1 as the increased attendance group because attendance on the day of the test is higher for all GPA deciles compared to predicted attendance. We label cluster 2 as strategic schools because of the sizable difference between observed and predicted attendance for lower-performing schools, which averages -18 and -16 percentage points for the first and second decile in terms of GPA. Importantly, both groups of schools have similar attendance patterns before the test, in terms of average attendance, but also attendance by GPA decile.

Figure 7: Predicted vs. Observed Attendance on the Day of the Test
(4th Grade 2017, RM Region)



Notes. This figure illustrates that while our prediction model accurately predicts attendance for non-tested grades, lower-performing students attend less than predicted, while students who are not at the bottom of the GPA distribution attend more than predicted. These patterns are consistent with our event-study results. The graph is constructed using daily attendance data from publicly funded schools, with 4th grade in 2017 as the tested grade and 1st, 2nd, 3rd, 5th, and 6th grades as non-tested grades. Each dot represents the mean of predicted or observed attendance at the GPA-decile–school–day level in the Metropolitan Region of Chile. Unlike the event-study estimates, school-level data allow us to develop a strategy to identify strategic and non-strategic schools.

Table 7 also reports the average characteristics of both clusters, based on their overall performance and socioeconomic characteristics. The increased attendance group perform better in terms of SIMCE Math score,s with 5 percentage point less of students having below basic scores and 4 percentage point more in terms of proficient students[12]. Importantly, they are three times more likely to be classified as Insufficient in the school accountability measured based on academic and non-academic indicators of the past three years (called *Cateogrías de Desmepeño*) measured by the Quality Education Agency that can drive a school into

[12]Students are classified into three groups based on their mathematics scores: below 245 points, between 245 and 295 points, and above 295 points. These educational standards and score thresholds are defined by the Ministry of Education and approved by the National Council of Education. Similar differences between groups are observed in other subjects. This classification are essential for accountability, in particular for the school classification presented below.

closure.

Table 7: Characteristics of Clusters Defined by Differences Between Observed and Predicted Attendance

	Increase Att. (1) (N = 577)		Strategic Att. (2) (N = 157)		(1) - (2)	
	Mean	SD	Mean	SD	Diff.	p-value
Observed–Predicted Attendance Difference on Test Day						
Diff D1 GPA	0.01	(0.10)	-0.17	(0.20)	-0.18***	0.00
Diff D2 GPA	0.03	(0.07)	-0.13	(0.14)	-0.16***	0.00
Diff D3D8 GPA	0.03	(0.03)	-0.04	(0.07)	-0.07***	0.00
Diff D9 GPA	0.03	(0.05)	-0.01	(0.08)	-0.05***	0.00
Diff D10 GPA	0.02	(0.05)	0.01	(0.07)	-0.01*	0.09
Attendance						
Attendance (Before Test Day)	0.91	(0.03)	0.91	(0.03)	-0.00	0.86
Attendance (Test Day)	0.96	(0.03)	0.87	(0.06)	-0.08***	0.00
School Characteristics						
Priority Students	0.44	(0.19)	0.48	(0.19)	0.04*	0.02
SIMCE Score Below Basic (Math)	0.35	(0.20)	0.40	(0.19)	0.05**	0.01
SIMCE Score Basic (Math)	0.41	(0.10)	0.40	(0.10)	-0.01	0.17
SIMCE Score Proficient (Math)	0.24	(0.16)	0.20	(0.16)	-0.04*	0.01
High (Perf. Category)	0.16	(0.37)	0.08	(0.28)	-0.08**	0.00
Middle (Perf. Category)	0.53	(0.50)	0.44	(0.50)	-0.09*	0.04
Mid-low (Perf. Category)	0.25	(0.43)	0.31	(0.46)	0.06	0.12
Insufficient (Perf. Category)	0.06	(0.24)	0.17	(0.37)	0.11***	0.00

Notes. (a) This table compares the observable characteristics of two k-means optimal clusters defined by their differences between observed and predicted attendance. The table shows that the second cluster exhibits characteristics consistent with strategic behavior, despite having SES characteristics similar to those of the first cluster. (b) 'Diff DX GPA' represents the difference between obs. attendance and predicted attendance for decile X. Increase Att. (1) represents Cluster 1, where overall attendance increases for all GPA deciles. Strategic Att (2) represents Cluster 2, consistent with strategic behavior. (c) Priority Students are students defined as lower income by the SEP targeted voucher policy. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Penalization Strategy Considering Strategic Behavior

	Increase Att. (1) (N = 577)		Strategic Att. (2) (N = 157)		(1) - (2)
	Mean	SD	Mean	SD	Diff.
Attendance < 95%	0.43	(0.50)	0.98	(0.14)	0.55***
SIMCE Score Below Basic					
No Penalization	0.35	(0.20)	0.40	(0.19)	0.05**
Only Strategic Schools	0.35	(0.20)	0.49	(0.21)	0.13***
All Schools	0.37	(0.21)	0.49	(0.21)	0.12***

Notes. (a) This table shows that 43% of schools in the "increase attendance" group that do not appear to behave strategically would be affected by a simple penalization strategy applied to any school with less than 95% attendance on the test day, similar to those implemented in nearly 24 U.S. states following ESSA rules Katanyoutanant et al. (2021). At the same time, almost all schools labeled as strategic would be penalized. (b) * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Finally, Table 8 shows how many schools would be affected by a simple penalization strategy applied to any school with less than 95% attendance on the test day, similar to those implemented in nearly 24 U.S. states (Katanyoutanant et al., 2021). Falling below the 95% participation threshold can reduce accountability scores (e.g., Hawaii, New Mexico, Ohio, Oklahoma, Rhode Island, Tennessee, Washington), trigger escalating penalties (e.g., Maryland), or require schools to submit corrective action plans (e.g., Florida), among many other possibilities (Katanyoutanant et al., 2021). We find that 43% of schools not classified as strategic would be subject to such penalties, representing roughly one-third of all schools. However, these non-strategic schools would be substantially less affected by penalties that impute missing test-takers as below basic, as their students exhibit higher SIMCE scores. At the same time, almost all of the schools labeled as strategic would be penalized.

7 Conclusion

High-stakes testing is widely used across educational systems, shaping accountability, resource allocation, and school choice. Because skewed attendance may undermine these goals, it is essential to understand the impact of testing on attendance beyond the exclusion of low-performing students. Using an extensive national panel of daily attendance, this article presents causal evidence that, in a country with a long tradition of high-stakes testing, test-day exclusion is only part of the story, challenging the traditional narrative that focuses on the exclusion (or self-exclusion) of low-performing students (Coelli and Foster, 2024; Cullen and Reback, 2006; Cilliers et al., 2019; Feigenberg et al., 2019; Figlio and Getzler, 2002; Figlio, 2006; Hofflinger and Hippel, 2018; Hofflinger and Hippel, 2020; Jacob, 2005).

We find negative attendance effects only among younger, low-performing students and positive effects for students above the bottom of the GPA distribution across all grades. These compositional shifts are largely masked in aggregate estimates, which typically suggest small positive attendance effects and align with how policymakers measure attendance. As a result, test scores may reflect a selected test-taking population, with implications for accountability and resource allocation, particularly among younger students. Using survey data, a separate event study, and exemption records, we rule out student-exemption policies and test anxiety as mechanisms. Instead, the evidence is consistent with two channels beyond the traditional exclusion of low-performing students: informational advantages among higher-performing students, reflected in stronger student–parent–school communication, and the salience of rewards and grading incentives among younger, low-performing students.

Our study has several policy implications for the design of accountability systems in both developed and developing countries. We highlight the importance of monitoring and reporting test-taker attendance by pre-test performance levels, such as GPA. Even after three decades of high-stakes testing, aggregate attendance statistics have been presented as evidence that Chilean institutions perform well on test-day attendance. However, heterogeneity in the impact of testing on attendance masks an urgent need to change how test-day attendance is reported, since aggregate measures can conceal non-representative participation patterns that differ across grades and are more urgent among younger students. This issue has remained largely unaddressed in national debates around SIMCE, despite persistent concerns about student exclusion Equipo de Tarea (2015). We also show that, although multiple imputation can reduce bias in expectation, school-level imputations may still be far from true values because of missing-data variability, especially in smaller schools. This strategy is used in Australia's NAPLAN and has been proposed for Chile by Cuesta et al. (2020b) . Similarly, across-the-board absence penalties—such as those implemented under the ESSA in the U.S.—may increase attendance but risk sanctioning schools whose test-day attendance falls below thresholds for reasons related to student composition rather than strategic behavior. To address this tension, we propose a data-driven prediction strategy that forecasts school-by-decile attendance on the test day using pre-test attendance dynamics and compares predicted with observed attendance. This approach identifies a subset of schools with unusually large negative deviations among lower-performing students, consistent with strategic behavior, and allows policymakers to target oversight and support rather than rely on blunt participation rules. Finally, independent of exclusion patterns, our findings highlight the need to improve communication among students, parents, and schools, particularly for younger and low-performing students. Distinguishing and addressing the positive and negative aspects of school strategies—such as prizes and awards—remains a central task for future work.

References

AGCE. 2015a. Síntesis de resultados de aprendizaje SIMCE 2014: 4° educación básica. Education Quality Agency: Santiago, Chile.

AGCE. 2015b. Síntesis de resultados de aprendizaje SIMCE 2014: 6° educación básica. Education Quality Agency: Santiago, Chile.

Chapell M, Blanding Z, Silverstein M, Takahashi M, Newman B, Gubi A, McCann N. 2005. Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology* **97**: 268–274.

Cilliers J, Mbiti I, Zeitlin A. 2019. Can public rankings improve school performance? *Journal of Human Resources*.

Coelli M, Foster G. 2024. Unintended consequences of school accountability reforms: Public versus private schools. *Economics of Education Review* **99**: 102523. Available from: https://doi.org/10.1016/j.econedurev.2024.102523.

Cuesta JI, González F, Larroulet C. 2020a. Distorted quality signals in school markets. *Journal of Development Economics*.

Cuesta JI, González F, Larroulet C. 2020b. Distorted quality signals in school markets.

Cullen J, Reback R. 2006. Tinkering toward accolades: School gaming under a performance accountability system. National Bureau of Economic Research, Inc. Available from: https://EconPapers.repec.org/RePEc:nbr:nberwo:12286.

Equipo de Tarea. 2015. Hacia un sistema completo y equilibrado de evaluación de los aprendizajes en chile: Informe equipo de tarea para la revisión del SIMCE. Ministerio de Educación, Gobierno de Chile. Available from: https://www.mineduc.cl/wp-content/uploads/sites/19/2015/11/Informe-Equipo-de-Tarea-Revisi%C3%B3n-Simce.pdf.

Feigenberg B, Yan R, Rivkin S. 2019. Illusory gains from chile's targeted school voucher experiment. *The Economic Journals* **54**: 2805–2832.

Figlio D. 2006. Testing, crime and punishment. *Journal of Public Economics* **90**(4-5): 837–851. Available from: https://ideas.repec.org/a/eee/pubeco/v90y2006i4-5p837-851.html.

Figlio D, Getzler L. 2002. Accountability, ability and disability: Gaming the system. National Bureau of Economic Research, Inc.

Freyaldenhoven S, Hansen C, Shapiro JM. 2019. Pre-event trends in the panel event-study design. *American Economic Review* **109**(9): 3307–38. Available from: https://www.aeaweb.org/articles?id=10.1257/aer.20180609.

Hart R, Casserly M, Uzzell R, Palacios M, Corcoran A, Spurgeon L. 2015. Student testing in america's great city schools: An inventory and preliminary analysis. Council of the Great City Schools. Available from: https://www.cgcs.org.

Hembree R. 1988. Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research* **58**(1): 47–77. Available from:https://doi.org/10.3102/00346543058001047 .

Hofflinger A, Hippel P von. 2020. Missing children: How chilean schools evaded accountability by having low-performing students miss high-stakes tests. *Educational Assessment Evaluation and Accountability*. Available from: https://doi.org/10.2139/ssrn.2906552.

Hofflinger A, Hippel P von. 2018. The response to high-stakes testing in chile: Did schools improve learning or merely inflate test scores? *Available at SSRN 2906552*.

Jacob BA. 2005. Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of Public Economics* **89**(5): 761–796. Available from: https://doi.org/10.1016/j.jpubeco.2004.08.004.

Katanyoutanant T, Kwon S, McNeilly M, Hemberg B, Thurlow ML, Lazarus SS, Hinkle AR, Liu KK. 2021. The 95 percent state assessment participation requirement: Current landscape, state challenges, and recommended strategies (NCEO report 429). National Center on Educational Outcomes, University of Minnesota. Available from: https://nceo.umn.edu/docs/OnlinePubs/NCEOReport429.pdf.

Mercurio E. 2024. Advierten riesgos de eliminacion de prueba simce de lectura en 2 basico por rezago de escolares. *El Mercurio*. Available from: https://mcusercontent.com/b38458f930f0bf51d3800a14c/images/4958d76f-aaea-f26b-173b-ee484295421a.jpeg.

Neilson CA. 2025. Targeted vouchers, competition among schools and the academic achievement of poor students. *Econometrica (Conditionally Accepted)*. Available from: https://christopherneilson.github.io//work/documents/Neilson_JMP/Neilson_SEPVouchers_April_2025.pdf.

OCDE. 2013. *Synergies for Better Learning*. Available from: https://doi.org/10.1787/9789264190658-en.

OECD. 2023. *Education at a Glance 2023: OECD Indicators*. OECD Publishing: Paris. Available from: https://doi.org/10.1787/e13bef63-en.

OECD. 2017. *PISA 2015 Results (Volume III)*. Available from: https://www.oecd-ilibrary.org/content/p ublication/9789264273856-en.

Paladino M. 2020. Towards an understanding of the testing opt-out movement: Why parents choose to opt-out or opt-in. *Journal for Leadership and Instruction* **Fall**: 14–? Available from: https://files.eric.ed.gov/fulltext/EJ1283112.pdf.

SACLaw. 2011. Sistema nacional de aseguramiento de la calidad de la educación parvularia, básica y media y su fiscalización. (20.529). Available from: https://www.bcn.cl/leychile/navegar?idNorma=1028635.

Seipp B. 1991. Anxiety and academic performance: A meta-analysis of findings. *Anxiety Stress and Coping - ANXIETY STRESS COPING* **4**: 27–41.

Sigrist F. 2022a. Gaussian process boosting. *Journal of Machine Learning Research* **23**(232): 1–46.

Sigrist F. 2022b. Latent gaussian model boosting. *IEEE* **45**(2): 1894–1905.

SIMCE. 2006. *Informe de Resultados 4° Básico 2005.* Minsiterio de Educación de Chile.

SIMCE. 2007. *Resultados Para Docentes y Directivos 4° Básico 2006.* Minsiterio de Educación de Chile.

SIMCE. 2008. *Resultados Para Docentes y Directivos 8° Básico 2007.* Minsiterio de Educación de Chile.

U.S. DOE. 2017. *Accountability, State Plans, and Data Reporting: Summary of Final Regulations Under the Every Student Succeeds Act (ESSA).* U.S. Department of Education. Available from: https://www.ed.gov/sites/ed/files/policy/elsec/leg/essa/essafactsheet170103.pdf.

U.S. DOE. 2016. *Summary: Proposed Regulations on Accountability, State Plans, and Data Reporting Under the Every Student Succeeds Act (ESSA).* U.S. Department of Education. Available from: https://www.ed.gov/sites/ed/files/policy/elsec/leg/essa/essaaccountabilitynprmsummary52016.pdf.

Van Buuren S. 2018. *Flexible Imputation of Missing Data.* Second edition. Chapman & Hall/CRC. Available from: http://gen.lib.rus.ec/book/index.php?md5=5BEBB49E0EE338EB87B749B950825C82.

A Appendix

A.1 Figures

Figure A1: Examples of Pre-filled Letters Sent by the Government

Annex 1. Communication Template for Parents and Guardians Sent by the School Principal [ENGLISH TRANSLATION]

[City, date]

Dear Parent or Guardian:

On the days ______________, students in 5th grade will take the SIMCE Experimental Test in Language and Communication, Mathematics, History, Geography, and Social Sciences. The purpose of this assessment is to study how a set of questions functions, in order to later select those that will be used to prepare the final tests. Therefore, these tests will not yield results for the school, but it is very important that all students participate, as this will allow the Ministry of Education to select the most appropriate questions that will be included in the final assessments to be taken by all 5th-grade students nationwide.

We kindly request your cooperation during these days by ensuring your child's attendance at school. It is advisable that the student has had a good night's rest beforehand, and that they arrive calm and ready to give their best effort. During the administration of the test, students are not allowed to bring or use electronic devices, such as cell phones, MP3 players, calculators, etc.

Therefore, we recommend that you ensure that your child does not bring any of these devices on that day.

Thanking you in advance for your cooperation,

Principal

Anexo 1: Modelo de comunicación para padres y apoderados que envía el director del establecimiento

Ciudad, fecha

Señor apoderado:

Los días , los alumnos de 5º básico rendirán la Prueba Experimental SIMCE, de Lenguaje y comunicación, Matemática e Historia, geografía y ciencias sociales. El propósito de esta aplicación es poder estudiar el funcionamiento de un conjunto de preguntas, para posteriormente seleccionar aquellas que permitan preparar las pruebas definitivas. Por lo tanto, estas pruebas no arrojarán un resultado para el establecimiento, pero sí es muy importante que todos los alumnos participen, puesto que esto permitirá al Ministerio de Educación seleccionar las mejores preguntas que serán incluidas en las pruebas definitivas que todos los estudiantes de 5º básico del país deberán rendir.

Solicitamos encarecidamente contar con su colaboración durante estos días, propiciando la asistencia de su hijo al establecimiento. Es conveniente que el alumno haya tenido un buen descanso la noche anterior, que se encuentre tranquilo y dispuesto a entregar su mejor esfuerzo.

Durante la aplicación de las pruebas no está permitido portar ni usar elementos electrónicos, tales como celulares, MP3, calculadoras, etc. Por ello, se recomienda que usted se asegure que este día el alumno no lleve consigo ninguno de estos aparatos.

Esperando contar con su cooperación, le saluda atentamente,

Director

Annex 2. Informational Communication (ENGLISH TRANSLATION)

[City, month and year]

Dear Parent or Guardian:

During the day(s) _________ of _________ (month), students in _________ grade (elementary/high school) will take the SIMCE assessments.

It is important that the student attends school on the day(s) the tests are administered, as an absence could mean that the school is unable to include their results.

It is advisable that the student has had a good night's rest the night before, and that they arrive calm and ready to give their best effort.

During the administration of the tests, students are not permitted to bring or use electronic devices, such as cell phones, audio players, calculators, etc. Therefore, we recommend that you ensure that on those days the student does not bring any of these devices.

Your child will be given the Parent and Guardian Questionnaire along with an envelope so that you may complete your information. Please respond and return it in the sealed envelope on the day _______________.

For your peace of mind, we inform you that all information provided in this questionnaire is strictly confidential and for the exclusive use of the Agency for Quality of Education.

Thanking you for your cooperation,

[Signature]
Principal

Anexo 2. Comunicación informativa

[ciudad, mes y año]

Señor(a) apoderado(a):

Durante el (los) día(s) .... de .... (mes) los estudiantes de .... básico/medio rendirán las pruebas Simce.

Es importante que el estudiante asista al establecimiento durante el (los) día(s) de aplicación de las pruebas, debido a que una baja asistencia puede significar que nuestro establecimiento no pueda acceder a los resultados de ellas.

Es conveniente que el estudiante haya tenido un buen descanso la noche anterior, que se encuentre tranquilo y dispuesto a entregar su mejor esfuerzo.

Durante la aplicación de las pruebas no está permitido portar ni usar elementos electrónicos, tales como celulares, reproductores de audio, calculadoras, etc. Por ello, se recomienda que usted se asegure de que en esos días el estudiante no lleve consigo ninguno de estos aparatos.

Su estudiante le entregará el Cuestionario Padres y Apoderados y un sobre para que usted registre sus datos, lo responda y lo devuelva en el sobre sellado el día ....

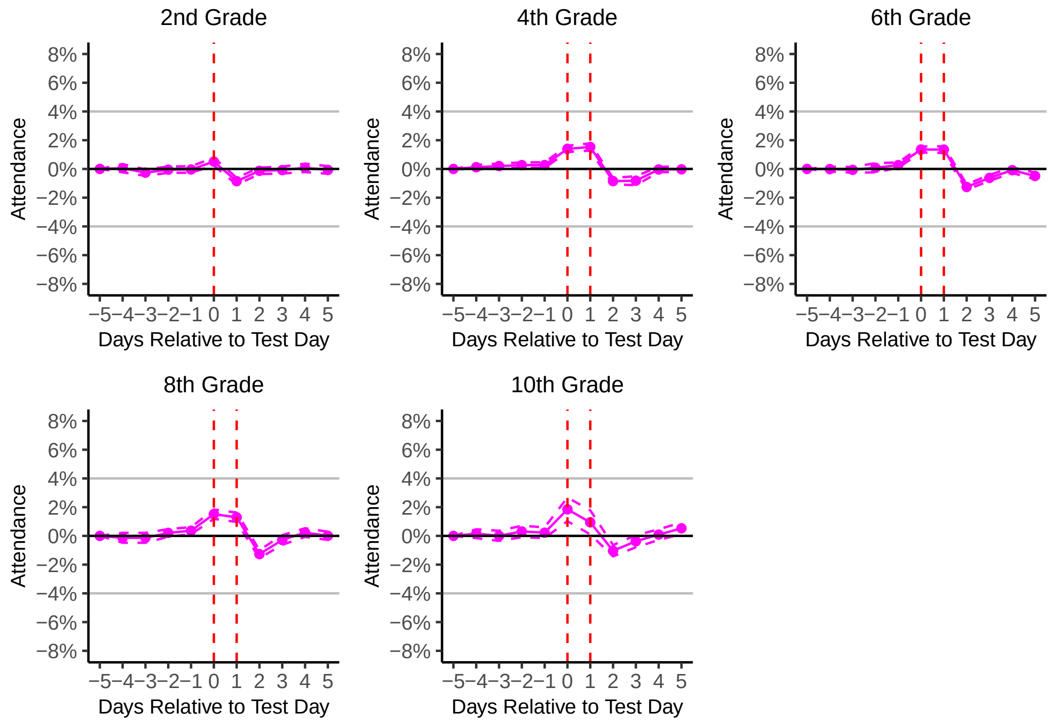
Para su tranquilidad, le comunicamos que toda la información entregada en dicho cuestionario es totalmente confidencial y de uso exclusivo de la Agencia de Calidad de la Educación.

Esperando contar con su cooperación, lo(a) saluda atentamente,

[firma]
Director

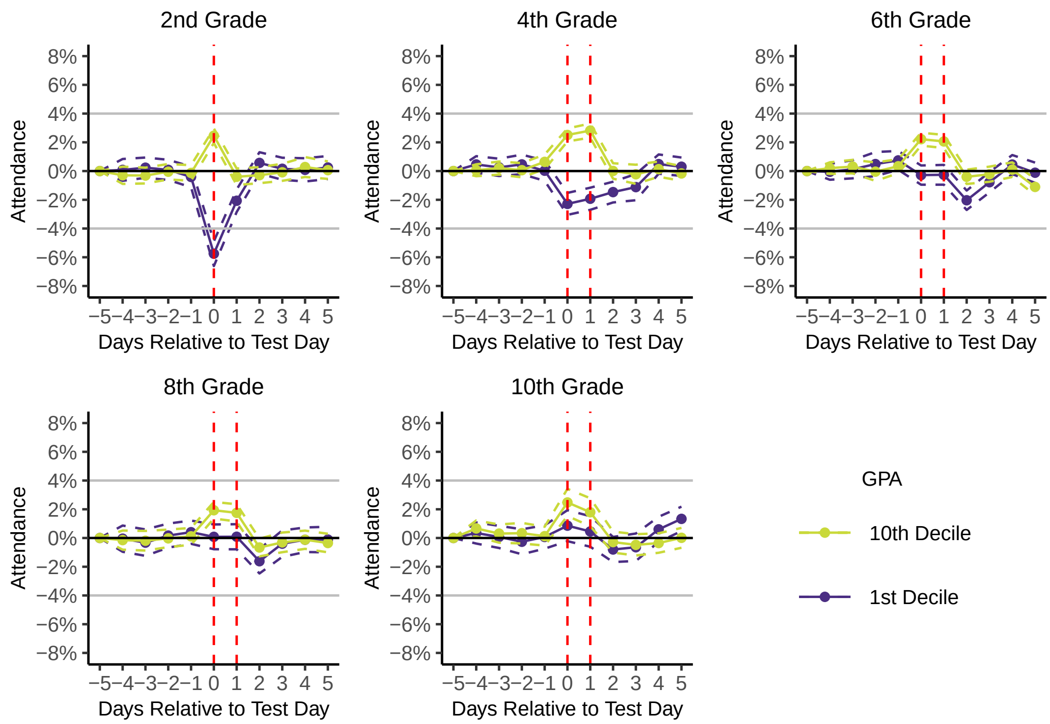
Notes. The image on top show the pre-filled letter sent by the government for all the SIMCE tests in 2017. The images on the bottom shows the pre-filled latter sent by the government for the 5th grade no-stakes test in 2012

Figure A2: Event Study Estimates Around the Test Day (2014, All Grades)



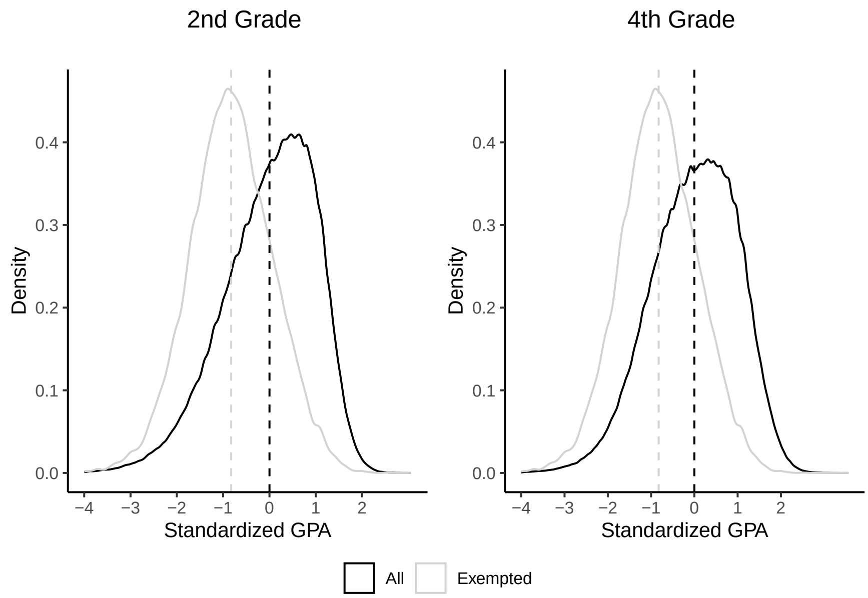
Notes. Read footnote for 2011.

Figure A3: Event Study Estimates Around the Test Day (2014, All Grades)
Low- and High-Performing Students



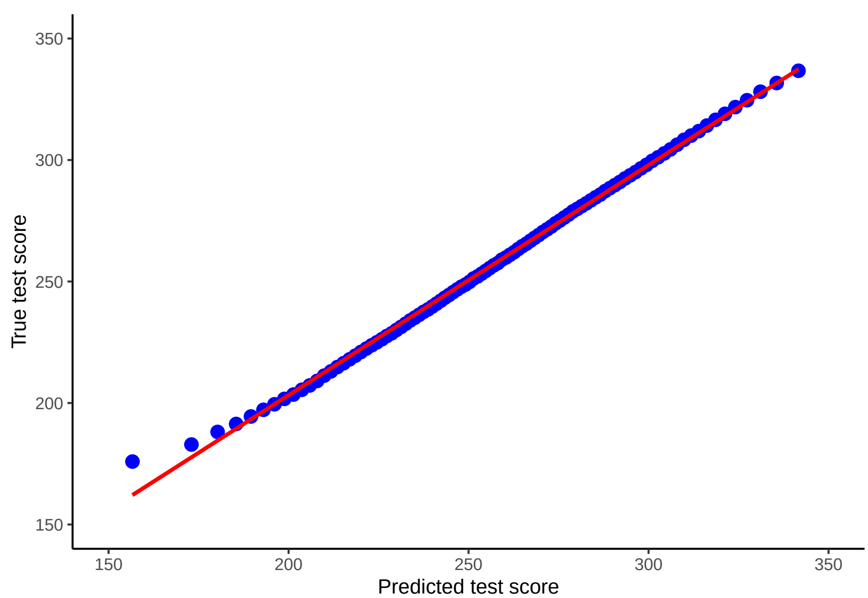
Notes. (1) This figure expands Figure 2 and illustrates no pre-trends supporting the parallel trends assumption for any grade, while also showing the attendance gap between grades taking and not-taking the test come back to pre-trend levels soon after the day of the test. (2) Dashed lines are 95% confidence intervals and the vertical red lines show the days in which the test is taken. In all grades except 2nd grade, the test was administered over two consecutive days.

Figure A4: Example of Standardized GPA Density for Exempted Students (2014)



Notes. (1) Dotted lines are the distribution means. (1) This figure illustrates that exempted students were low performers, suggesting that exemptions may represent a potential pathway for score manipulation.

Figure A5: Replication of Cuesta, et al (2020b) Figure A.11



A.2 Tables

Table A1: Impact of the Test Day on Attendance (2011, 2016-2017)

Year/Grade	D1	D2	D3D8	D9	D10	D10-D1	ALL
2011							
4th	-0.01***	0.00*	0.01***	0.02***	0.02***	0.03***	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
8th	0.01**	0.02***	0.03***	0.03***	0.03***	0.02***	0.03***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
2016							
4th	-0.04***	-0.02***	0.01***	0.02***	0.02***	0.06***	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
6th	-0.01***	0.00	0.01***	0.02***	0.02***	0.03***	0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
10th	-0.01	0.00	0.00	0.00	0.01**	0.01***	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
2017							
4th	-0.05***	-0.02***	0.00***	0.02***	0.02***	0.07***	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
8th	-0.02***	-0.01**	0.00*	0.01**	0.01***	0.03***	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)
10th	-0.02***	-0.02***	0.00**	0.00	0.00	0.02***	-0.01***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)

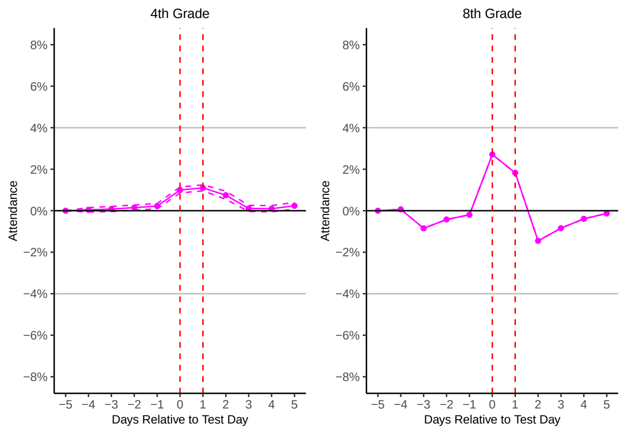
Notes. (1) This table shows the main results from Equation 2, except for the last columns, which present results from the event-study specification that does not account for heterogeneity presented in Equation 4.1. (2) D1, D2, D9, and D10 correspond to the first, second, ninth, and tenth GPA deciles, respectively. D3D8 corresponds to the third through eighth GPA deciles. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the school level.

B Online Appendix

B.1 Online Appendix Figures

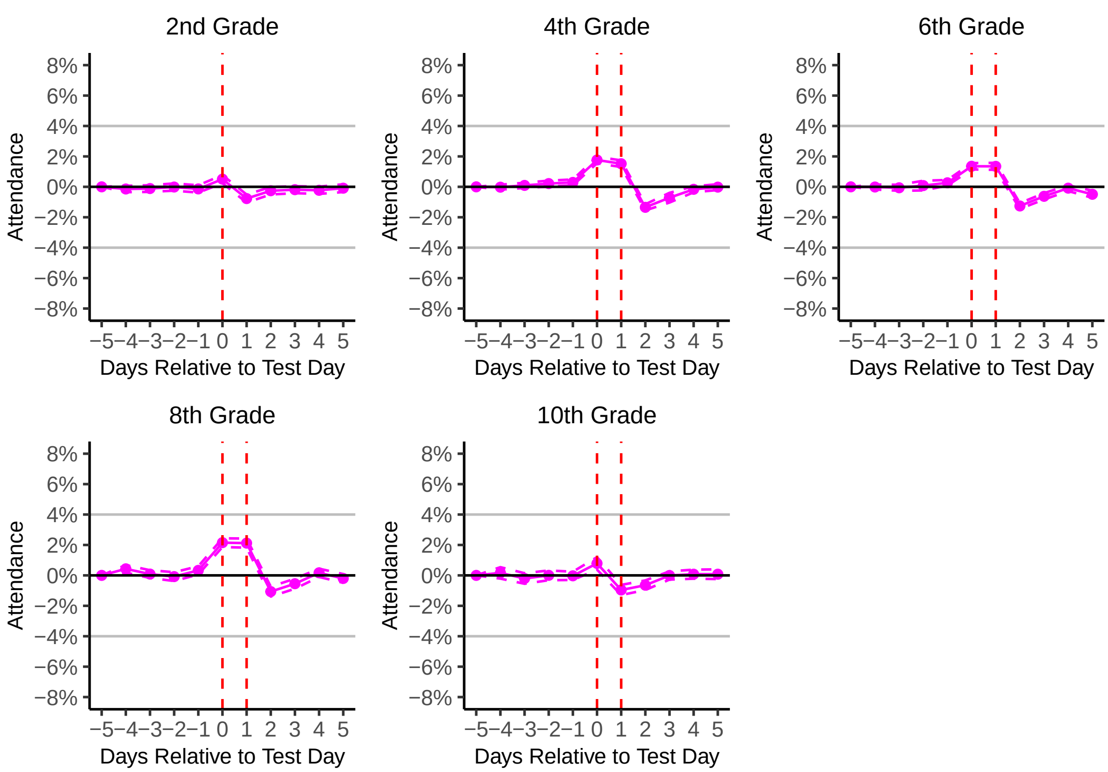
B.1.1 Event Study Estimates Around the Test day

Figure B1: Event Study Estimates Around the Test Day (2011, All)



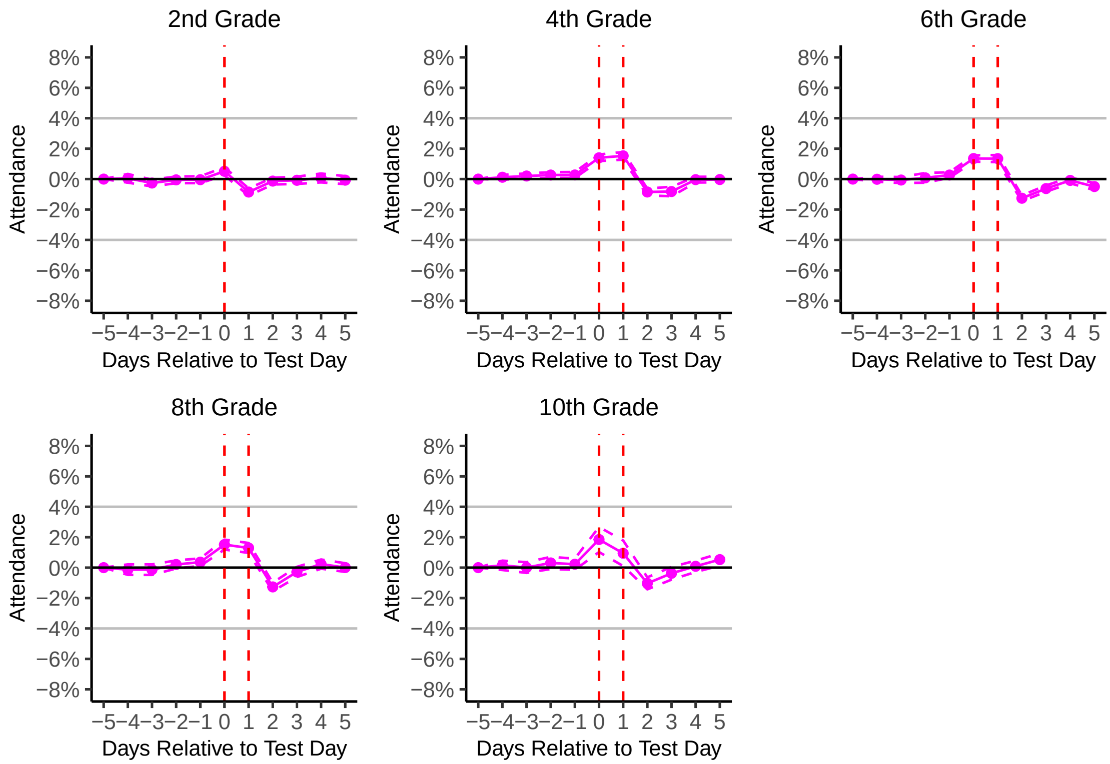
Notes. (1) This figure expands illustrates no pre-trends supporting the parallel trends assumption for any grade, while also showing the attendance gap between grades taking and not-taking the test come back to pre-trend levels soon after the day of the test. (2) Dashed lines are 95% confidence intervals and the vertical red lines show the days in which the test is taken. In all grades except 2nd grade, the test was administered over two consecutive days.

Figure B2: Event Study Estimates Around the Test Day (2013, All)



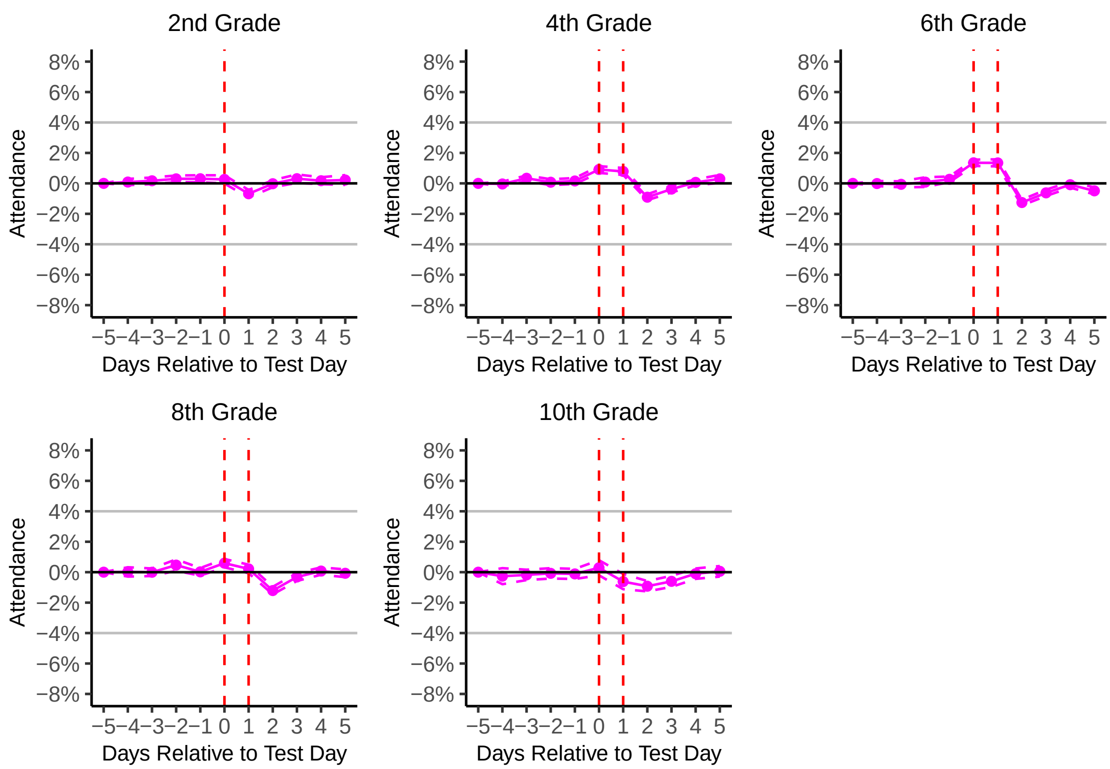
Notes. Read footnote for 2011.

Figure B3: Event Study Estimates Around the Test Day (2014, All)



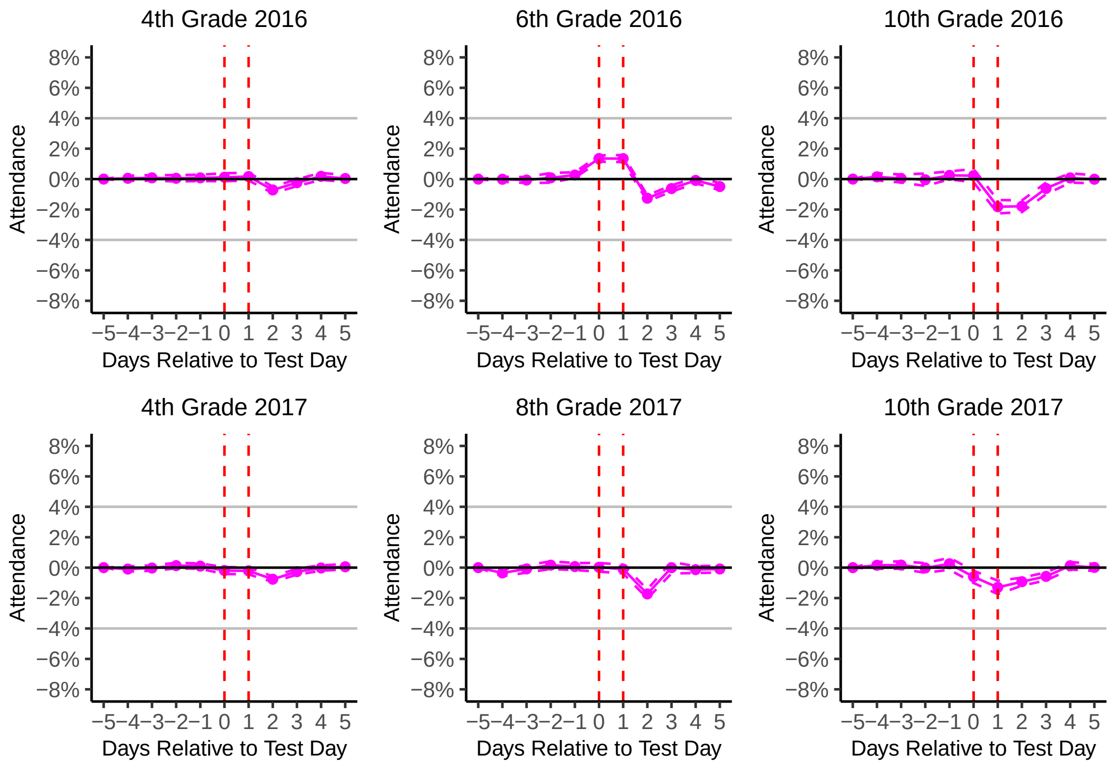
Notes. Read footnote for 2011.

Figure B4: Event Study Estimates Around the Test Day (2015, All)



Notes. Read footnote for 2011.

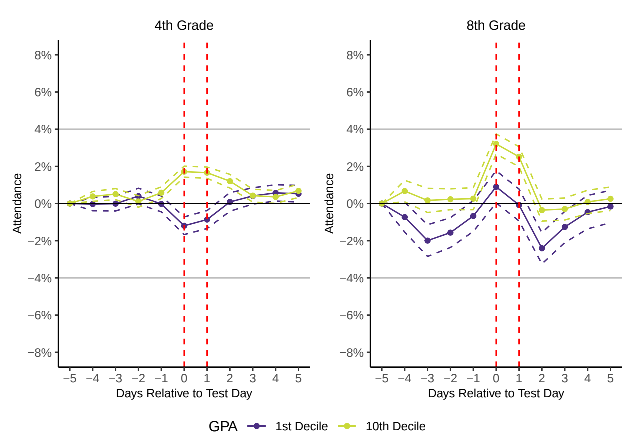
Figure B5: Event Study Estimates Around the Test Day (2016, 2017, All)



Notes. Read footnote for 2011.

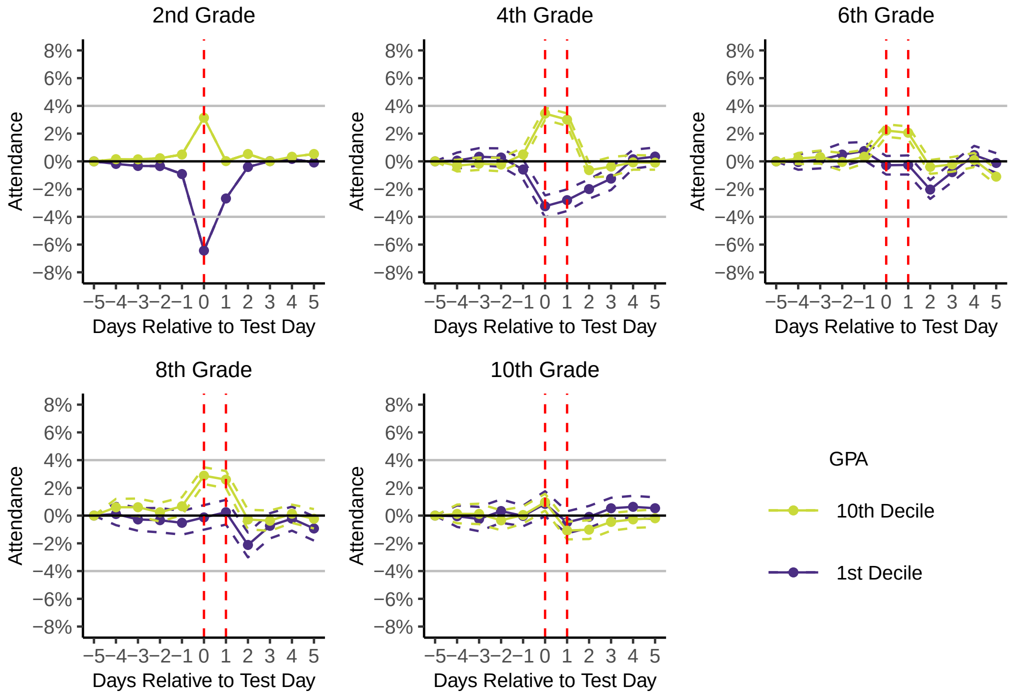
B.1.2 Event Study Estimates Around the Test day (Low- and High-Performing Students)

Figure B6: Event Study Estimates Around the Test Day (2011, All Grades)
Low- and High-Performing Students



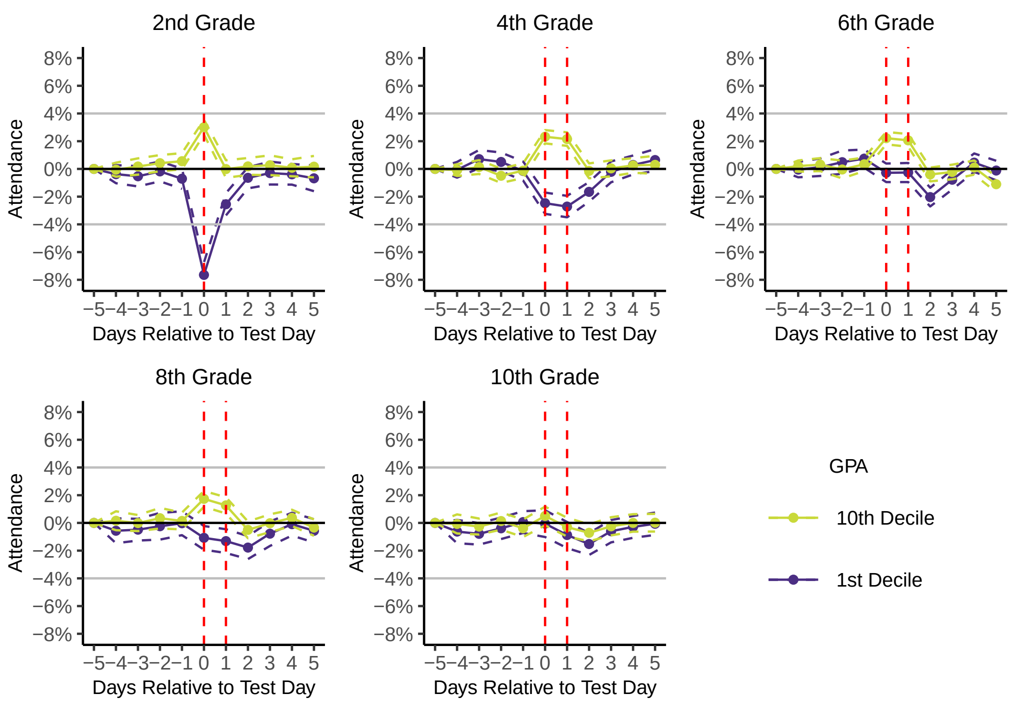
Notes. (1) This figure expands Figure 2 and illustrates no pre-trends supporting the parallel trends assumption for any grade, while also showing the attendance gap between grades taking and not-taking the test come back to pre-trend levels soon after the day of the test. (2) Dashed lines are 95% confidence intervals and the vertical red lines show the days in which the test is taken. In all grades except 2nd grade, the test was administered over two consecutive days.

Figure B7: Event Study Estimates Around the Test Day (2013, All Grades)
Low- and High-Performing Students



Notes. Read footnote for 2011.

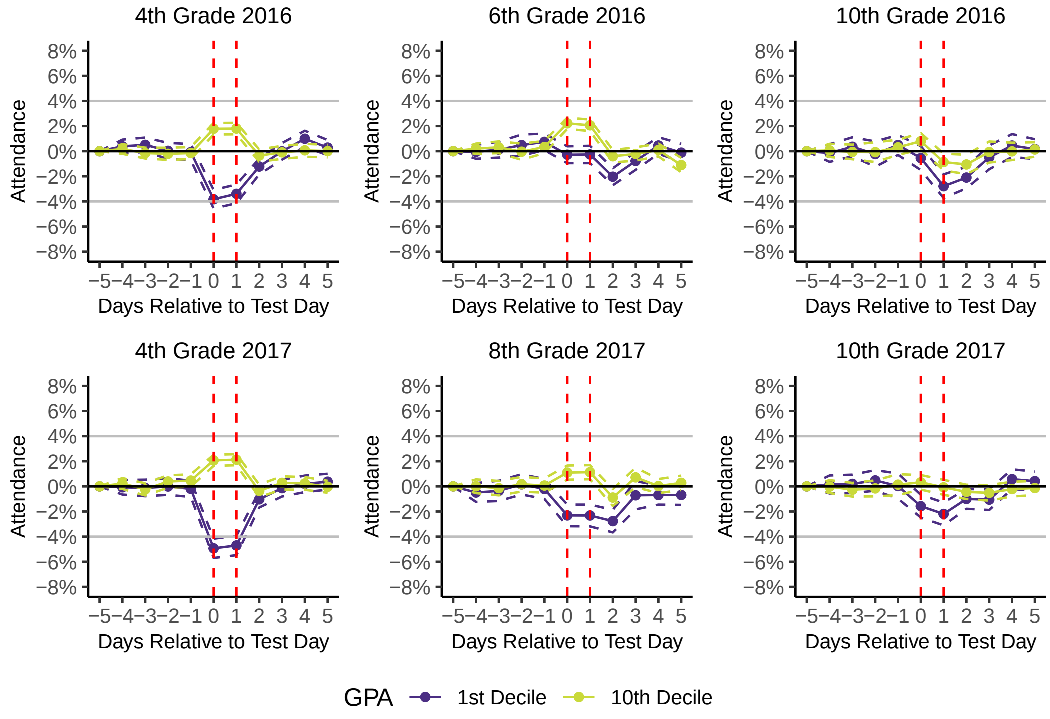
Figure B8: Event Study Estimates Around the Test Day (2015, All Grades)
Low- and High-Performing Students



Notes. Read footnote for 2011.

Figure B9: Event Study Estimates Around the Test Day (2016 - 2017, All Grades)

Low- and High-Performing Students



Notes. Read footnote for 2011.