Heterogeneous Incentives to Distort Performance Metrics: Evidence from a Large-scale RCT *

Micole De Vera[†] Javier Garcia-Brazales[‡] Jiayi Lin[§]

November 2026

Abstract

Ensuring standardized performance metrics are implemented consistently is essential for extracting reliable quality signals used in resource allocation. When agents face heterogeneous incentives to perform well, such metrics may induce strategic behavior and distort comparisons. Exploiting random assignment to external versus internal evaluation in primary schools in Madrid, we show that external evaluation lowers average test scores. Crucially, this effect is driven by private and charter schools, which face stronger incentives to appear high-performing. Their scores fall by approximately 0.30 standard deviations more than those of public schools. Distributional analyses reveal that internal evaluation shifts non-public schools from the middle of the score distribution into its upper half, so that they are over-represented at the top of school rankings. Under external evaluation, their share of the top decile would fall from about 55% to between 16% and 38%. Overall, our findings suggest that heterogeneous performance incentives can lead to systematic misranking that raises equity concerns when metrics are used for high-stakes decisions.

Keywords: Standardized exams; External and Internal Grading; Score Manipulation; School Rankings; RCT.

*This is a substantially revised version of a previously circulated paper entitled "May Internally-evaluated Standardized Exams Increase Cross-school Achievement Inequality? Evidence from a Randomized Controlled Trial." We thank Manuel Arellano, Samuel Bentolila, Caterina Calsamiglia, Mónica Martínez-Bravo, and seminar participants at the Banco de España for their useful comments. The willingness of Consejería de Educación e Investigación of the Comunidad de Madrid to provide access to the restricted-use data employed in this paper is very much appreciated. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Banco de España or the Eurosystem.

[†]Corresponding author: Banco de España, IZA@LISER, and RFBerlin. micole.devera@bde.es

[‡]University of Exeter and LEAP (Bocconi). For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

[§]Universitat Pompeu Fabra

1 Introduction

Performance metrics such as test scores, hospital ratings, or productivity benchmarks are increasingly used to monitor agents and reward effort (Lavy, 2009; Chen and Sivey, 2021). However, when the agents that are evaluated care about the metric and can control aspects of their measurement, classic moral hazard problems arise: agents may strategically manipulate the metric to appear higher performing (Zou, 2021; Chen and Dillender, 2025). In many settings, the strength of incentives to manipulate these metrics varies systematically across agents. This asymmetric moral hazard distorts the relative ranking implied by these performance measures, leading to misallocation of resources and ultimately reducing social welfare.

In this paper, we examine such dynamics in the context of standardized diagnostic testing in Spain. Although these tests are designed to measure student learning, these assessments may also function as salient quality signals for families, administrators, and policymakers (Hastings and Weinstein, 2008; Greaves and Hussain, 2026). The Spanish system includes public, charter, and private schools that differ in autonomy, funding models, and exposure to competition. Public schools have more stable funding and face relatively weaker market pressure. In contrast, non-public (charter and private) schools rely more directly on enrollment and reputation, giving them stronger incentives to generate favorable performance signals. These institutional differences therefore shape the incentives to appear high-performing.

To formalize these views and derive theoretical predictions, we start by developing a stylized model in which schools strategically manipulate reported performance under internal evaluation, showing that stronger incentives among non-public schools can distort true quality rankings: non-public schools rank higher than their quality warrants and are over-represented at the top of the ranking. The model further predicts that introducing external monitoring mitigates these distortions, lowering both the scores and the ranking of non-public schools relative to public schools.

We empirically test the model's key implications in a close-to-ideal scenario. We exploit a large-scale randomized controlled trial (RCT) conducted in the Madrid region during the 2016–2017 academic year in which schools were randomly assigned to either external or internal monitoring and marking of standardized exams. Since we observe information for the universe of schools, both public and non-public, we can test whether external evaluation deters grade inflation by schools and which type of school sees its performance decline the most under external

evaluation.

The experiment covers the standardized examinations taken in the sixth grade of primary school (the LOMCE exams, named after the educational law that introduced them), which evaluates students in Spanish, English, mathematics, and technology. Under internal evaluation, both test supervision and grading are carried out by teachers within the school, with only the identity of the student remaining anonymous to the grader. Under external evaluation, these tasks are performed by independent personnel, and the identity of both the student and the school remain anonymous to the grader. We therefore take the scores under external evaluation as true signals of quality. Random assignment at the school level—validated through balance tests—enables clean comparisons of score distributions across testing regimes, thereby overcoming the identification challenges that typically arise in quasi-experimental settings.

We find that externally evaluated schools perform systematically worse than internally evaluated ones, with differences that are statistically significant and of around 10% of a standard deviation for the composite mark across the four subjects evaluated. At a more disaggregated level, the effects are significant for Spanish, mathematics, and technology, with the gap not being statistically significant for English. Although these differences could arise from multiple sources, these baseline results are consistent with the presence of (broadly defined) score manipulation under the weaker supervision of the internally-evaluated regime.

We continue by exploring whether the differential incentives for score manipulation between public and non-public schools are reflected in the data. If the poorer performance of externally-evaluated schools is indeed driven by score manipulation, we should find that non-public schools are the ones whose performance suffers the most under external evaluation. This is precisely what the data shows: while the distribution of scores for public schools displays modest differences between the internal and external regimes, sizeable differences arise among private and charter schools. In particular, for the composite score, external evaluation lowers the scores of non-public schools by around 30% of a standard deviation more than those of public schools.

Our results are robust to multiple checks. The most important concern is selective participation, especially if it is driven by students or their families. External evaluation increases non-participation in non-public schools relative to public schools, so the differential decline in scores could reflect who sits the exam rather than how it is evaluated. This threat is not problematic if driven by schools' decisions, for two reasons. First, if schools decide to keep weaker students from sitting the exam, this is part of the strategic behavior of schools that our study

aims to capture. Second, because absences rise under external evaluation, it biases our estimates toward zero. If instead students or their families decide not to sit the exam, this is not manipulation by schools. Using Lee (2009) bounds, which allow for differential selection into test-taking, we find that the difference in effects between non-public and public schools remains negative, between 18 and 34 points for the composite score (0.18 to 0.34 standard deviations, as scores are normalized to a standard deviation of 100), so selection cannot fully explain our results. To address any remaining differences in underlying quality across schools, we additionally control for each school's score in the previous year, when all schools were externally evaluated, and the gradient remains large and statistically significant, at around 29 points. Since relatively few non-public schools were externally evaluated, we also verify that our conclusions do not rely on large-sample approximations using randomization inference. Moreover, because our baseline specification weights all schools equally, we re-estimate it at the student level and obtain a gradient of around 31 points. Finally, a separate randomization of grade 10 exams provides complementary evidence of the same gradient, although the score records for that grade are substantially less complete.

We then investigate how score manipulation shapes the distribution of scores across schools and, in turn, school rankings. Among non-public schools, the effect of external evaluation is largest in the upper half of the distribution: 17 points at the 10th percentile, 30 at the median, and 46 at the 90th. Counterfactual distributions based on DiNardo et al. (1996) show that internal evaluation moves mass out of the middle of the non-public distribution and into its upper half. Among public schools, the effects are small throughout. These shifts translate into large changes in rankings for non-public schools, as the model predicts. External evaluation lowers the rank of non-public schools by 16 percentile points. Under external grading, their share of the top decile would fall from about 55% to between 16% and 38%, depending on whether the counterfactual lets the effect of internal evaluation vary along the score distribution. Even when the effect is held constant across schools, about a fifth of the schools in the reported top decile would not be there.

Contributions to related literature. Using a policy that randomly assigned schools to different monitoring regimes across the universe of public and non-public schools, we show that schools with stronger incentives to appear high-performing inflate their scores more, and that this asymmetry distorts school rankings. We highlight two main contributions to the literature.

First, we show that the effects of monitoring are heterogeneous across agents with different incentives to appear high-performing. A growing literature documents that agents respond strategically to monitoring and improve measured performance when they are watched, in pollution (Zou, 2021), health care (Chen and Dillender, 2025), and education (Jacob and Levitt, 2003; Bertoni et al., 2013; Borcan et al., 2017; Martinelli et al., 2018; Dee et al., 2019). Less is known about how responses differ across types of agents that operate under common rules but face different incentives. The closest evidence is Tyrefors Hinnerich and Vlachos (2017), who use blind external re-marks of Swedish national tests to show that voucher schools mark more leniently than municipal schools, in a setting where the tests inform the grades used for university admission. Because evaluation regimes in our setting were randomly assigned before the exam, we instead identify the total effect of internal evaluation, capturing responses through invigilation and participation as well as marking. We find that internal evaluation leaves the scores of public schools essentially unchanged but raises those of non-public schools, even in predominantly multiple-choice exams and even though the exams carry no academic consequences for students. This suggests that restricting marking discretion need not ensure comparability across sectors, and that schools' own reputational incentives are sufficient to distort measured performance.

Second, we show that this heterogeneity distorts rankings. School performance measures are noisy (Kane and Staiger, 2002), but families respond to published school performance (Hastings and Weinstein, 2008; Greaves and Hussain, 2026). Strategic responses to monitoring are often thought to bias the level of performance measures while leaving them informative about relative quality: Chen and Dillender (2025), for instance, find that nursing home inspection ratings still predict relative quality despite strategic responses around inspections. We show that this need not hold when incentives to manipulate differ across agents. Because we observe the universe of public, charter, and private schools in the experimental districts, we can construct counterfactual rankings that families and administrators would face under each evaluation regime. Internal evaluation moves non-public schools out of the middle of the score distribution and into its upper half, so that they are substantially over-represented at the top of the ranking, the part that is likely most salient for families and administrators. This is important because rankings guide school choice and, potentially, the allocation of public resources.

Outline of the paper. The remainder of the paper is organized as follows. Section 2 describes the institutional context and the experimental design. Section 3 describes the data and our sample selection. Section 4 presents our conceptual and empirical frameworks. Section 5 documents the differential responses of public and non-public schools to external evaluation and demonstrates their robustness. Section 6 examines the distributional effects of external evaluation and their implications for school rankings. Section 7 concludes. In the Appendix, we report additional figures and tables, additional results, and key derivations.

2 Institutional Context and Experimental Design

2.1 Institutional Context: LOMCE Exams, School Types, and Incentives for High Performance

Diagnostic testing under LOMCE. On 9 December 2013, Spain enacted the *Ley Orgánica para la Mejora de la Calidad Educativa* (LOMCE).[1] As stated in its preamble, the reform pursued two main objectives: reducing early school dropout—which at the time stood at approximately 28 percent, nearly twice the European average—and improving educational outcomes as measured by internationally comparable indicators, such as student performance on large-scale assessments including PISA. To achieve these goals, the reform rested on three core pillars: (i) increased school autonomy, (ii) greater decision-making authority for school principals, and (iii) the use of *external evaluations to monitor academic performance*. Within this framework, the LOMCE introduced universal external assessments in grades 3, 6, and 10.

Incentives to appear high-performing. These diagnostic exams were explicitly designed as low-stakes tests for students: results did not affect grade promotion or access to subsequent educational tracks. Instead, the stated objective was to provide students, parents, teachers, and schools with standardized benchmarks of academic performance.

Despite their formally low-stakes nature for students, these evaluations generated potentially important incentives for schools and teachers. The score reports distributed to households enabled detailed relative performance comparisons at multiple levels. In particular, parents received individualized score sheets comparing: (i) their child's performance to that of other students in the same class; (ii) the performance of their child's class to that of other classes

[1] Organic Law for the Improvement of Educational Quality. The LOMCE replaced the *Ley Orgánica de Educación* (LOE), approved on 3 May 2006. In 2020, several years after our study period, the LOMCE was replaced by a new educational law: the LOMLOE.

within the same school; and (iii) the performance of their child's school to that of other schools in the district and in the whole of Madrid. School principals were also provided with analogous information. Figure A.1 presents an example of an individual student's report card depicting these dimensions of comparison. Moreover, there are online aggregator websites that include this information for a subset of schools for easier comparisons.[2]

The granularity of the score reports and their dissemination across stakeholders creates incentives for both school administrators and teachers, even though exams carried no direct consequences to students. For administrators, the results provided a potential source of reputational accountability. The regulatory framework required education authorities to communicate evaluation results to the educational community using common indicators and taking socioeconomic and sociocultural context into account.[3] The Department of Education of the Community of Madrid also publicly released aggregate statistics, including district-level averages by school ownership (public, charter, and private). These data allowed parents to compare performance across areas and school types, potentially influencing school choice and giving administrators an incentive to improve reported results. The scope for such responses became particularly salient following the 2013 school choice reform, which increased student mobility across districts by weakening the link between admission probabilities and residential proximity to schools (Gortázar et al., 2023).

For teachers, comparisons across classes within the same school provided a common metric of relative performance. School administrators could use this information to assess teachers, while teachers could use it to signal their effectiveness within the school. The reports thus created a potential source of internal accountability alongside the reputational incentives faced by school administrators.

Differential incentives by school ownership. The strength of these incentives likely differed across school sectors. The Spanish education system comprises three types of schools, classified according to governance structure and sources of funding. Public schools are fully funded and managed by the state, whereas private schools are entirely privately funded and operated. Charter schools ("*centros concertados*" in Spanish) occupy an intermediate position: they are privately owned but largely publicly funded and, in principle, tuition-free for

[2] See, for example, educa2.madrid.org.

[3] Article 8.4 of *Real Decreto 1058/2015, de 20 de noviembre*, required evaluation results to be communicated to the educational community through common indicators for all Spanish schools, taking socioeconomic and sociocultural context into account. These indicators were subsequently defined by the *Resolución de 30 de marzo de 2016, de la Secretaría de Estado de Educación, Formación Profesional y Universidades*.

households, although they are legally permitted to request voluntary parental contributions. In practice, such payments are widespread and often take the form of fixed monthly fees.[4]

In terms of student admissions, charter schools were originally conceived as a mechanism to alleviate capacity constraints in public provision and are formally required, like public schools, to accept all applicants. Nonetheless, charter schools are disproportionately concentrated in attractive urban areas, whereas public schools account for approximately 90 percent of schools in municipalities with fewer than 10,000 inhabitants. Private schools, by contrast, retain full discretion over student admissions.

Beyond differences in ownership and funding, school types also differ markedly in the working conditions and incentives faced by teachers. In public schools, teacher salaries and job stability are centrally determined and independent of student performance, being tied instead to tenure and formal qualifications. In charter and private schools, principals have greater autonomy over hiring and firing decisions, resulting in lower job security for teachers and, in many cases, compensation schemes that are more closely linked to school performance.

Taken together, these institutional features imply that, even in assessments that are formally low-stakes for students, schools and teachers face markedly different incentives to perform well. Public schools operate under comparatively weak performance incentives, as teacher compensation and job stability are largely insulated from student outcomes and enrolment demand. By contrast, charter and private schools are more directly exposed to competitive and reputational pressures, since attracting and retaining students is more closely tied to perceived school quality. In addition, teachers in these schools typically face lower job security and greater managerial discretion over hiring and retention decisions.

Whether such incentives are strictly stronger in private than in charter schools is theoretically ambiguous, particularly when the costs of score manipulation are low. While private schools face greater market exposure and retain full discretion over admissions, charter schools may also rely heavily on performance signals to attract and retain fee-paying households despite operating within a publicly funded system. In our empirical analysis we will primarily pool charter and private schools together for interpretability and because the number of private schools is relatively small. In additional analyses, we also explore the differential effects for charter and private schools.

[4] Acknowledging this issue, a recent reform—the *Ley Orgánica de Modificación de la Ley Orgánica de Educación*—has sought to curb compulsory payments in compulsory education and to reinforce the voluntary nature of parental contributions. Current estimates place these payments at approximately €1,000 per year on average.

2.2 Research Design

The context of the paper is one in which standardized diagnostic testing has been ongoing for many years.[5] The school level scores were widely reported, including in commonly-used websites for school choice. These exams, and the 2016 year of the LOMCE were externally evaluated. This regime proved highly costly so the Department of Education decided to introduced the RCT.[6] As such, we consider external evaluation as the status quo, benchmark scenario. We also consider this as the no-manipulation benchmark: external regime features external evaluators designated by the DAT to go to the schools and the marking is done by another external evaluator with anonymized school and student codes. This likely rules out the possibility that external evaluators may have incentives to purposefully lower certain schools grades (e.g., damage non-public schools for ideological reasons) since they would not know which school they are evaluating.[7] We can therefore interpret experimental differences between comparable internal and external schools as deviations from true quality measures (as much as these tests can capture those).

We exploit a unique setting to causally quantify distortions in quality signals within a standardized evaluation system that allows scope for performance manipulation under asymmetric incentives. Specifically, we study the implementation of the LOMCE exams in the Community of Madrid, where the regional Department of Education introduced a randomized controlled trial (RCT) governing the evaluation procedures used in Grade 6 standardized assessments. Under the RCT, schools were randomly assigned to either internal or external evaluation procedures. This experimental variation allows us to examine whether the extent of performance manipulation across evaluation regimes (internal versus external) depends on school ownership types (public versus non-public).

The Community of Madrid is administratively divided into five territorial areas ("*Direcciones de Área Territorial*", DATs): Capital, North, South, East, and West, shown in Figure A.2. Throughout the paper, we refer to these territorial areas as districts. The regional Department of Education used this administrative division as the main stratification variable in the experimental design.

The experimental design followed a rotating leave-one-district-out scheme. In the 2016–2017

[5] The predecessor of the LOMCE exams were the *Pruebas de Conocimientos y Destrezas Indispensables* (Tests of Essential Knowledge and Skills), more commonly referred to by the initials CDI, conducted between 2005–2015.

[6] The total monetary cost associated with the LOMCE exams in 2017 was approximately 600,000 euros.

[7] One could still argue that external evaluators may choose to grade more harshly any test assigned to them. However, this could not explain why we observe differential responses across public and non-public schools. Moreover, the predominantly multiple-choice nature of the tests reduces even more the subjectivity in marking.

academic year (hereafter 2017), the Capital district was excluded from randomization and was instead fully externally evaluated. The remaining four districts constitute the experimental sample used in our analysis. Within these districts, schools were randomly assigned to either external or internal evaluation procedures. In subsequent academic years, a different district was excluded from randomization and fully externally evaluated, although data for other years are not available to us.

Treatment is defined as full external evaluation, under which both exam invigilation and grading were carried out either by personnel external to the school or by legally accredited teachers[8] with no affiliation to the evaluated school, selected by the corresponding DAT and coordinated by the school inspectors. Control schools, by contrast, conducted both invigilation and grading internally using teachers employed by the school who did not teach the tested grade (for grade 6 tests it was preferably grade 5 teachers). In both evaluation regimes, exam invigilation and grading were conducted by different individuals and grading was blind: answer sheets were anonymized using encrypted student identifiers, and graders did not observe students' personal information. Aside from the identity of invigilators and graders, all other testing conditions—including exam content, timing, and administration schedules—were identical across schools.[9]

Randomization was stratified along several dimensions. First, as mentioned, schools were stratified by district. The number of schools assigned to external evaluation within each district was chosen to ensure sufficient statistical power to estimate within-district treatment effects. Second, within each district, schools were further stratified by ownership, participation in the Comunidad de Madrid bilingual education program, and by school size (measured by total student enrollment). This ensures balance across treatment and control groups in terms of school scale, bilingual status and, importantly for our study, school ownership.[10]

We additionally observe data from the 2015–2016 academic year (hereafter 2016), prior to the introduction of the experimental variation. In that year, all Grade 6 exams were externally invigilated and blindly graded. Thus, whereas the 2016 assessments were uniformly externally evaluated across all schools, the 2017 exams introduced randomized variation between fully external and fully internal evaluation procedures, as illustrated in Table 1.

[8]These teachers were from the Guidance Department ("*Departamento de Orientación*") or from the Educational and Psychological Guidance Team ("*Equipo de Orientación Educativa y Psicopedagógica*").

[9]The implementation of the tests and the collection of answer sheets were managed by private firms specializing in outsourced services.

[10]The instructions governing the administration of the 2017 Grade 6 exams are set out in the *Resolución de 8 de mayo de 2017*, published in the *Boletín Oficial de la Comunidad de Madrid*, No. 110, on 10 May 2017.

Table 1: The RCT Design

Year	Supervision	Grading
2016	External Exam Supervision	External Grading
2017	Either fully internal or fully external exam	Same assignment as for supervision

3 Data and Sample Selection

3.1 Data

Scores. The LOMCE exams are designed to assess core academic subjects. In grade 6, English, Spanish, mathematics, and technology (science) are evaluated. Exams consist primarily of multiple choice questions, although the English and Spanish assessments also contain open-ended sections in which students are asked to write short responses based on a given prompt. We have access to administrative data on the scores for two academic years: 2016 and 2017. The data include (i) students' standardized test scores in all subjects, normalized to have a mean of 500 and a standard deviation of 100 prior to sample selection and including the Capital region, and (ii) students' complete set of responses to all exam questions, which are available only for 2017.

School Characteristics. For each school, we have access to administrative records on key characteristics: type of ownership (private, charter or public), the number of students in the school (broken down by grade), and whether the school is participating in the bilingual program.

Sociodemographic Characteristics. Sociodemographic characteristics come from two sources. First, the school administration reported, for each test taker, whether they are Spanish and their gender. Hence, this information is widely available. Second, the Madrid Government distributed questionnaires to all students and parents to collect relevant information on these two collectives. These surveys were voluntary, and as such their completion rate is not high: around 50% of students and families completed them. To verify that the external evaluation treatment was properly randomly assigned, we focus on comparing a set of meaningful variables that were collected both in the student and parental questionnaires, and as such we minimize the number of missing observations by obtaining information from either side. These variables are: time devoted by the child to doing homework, relatives' involvement in the child's homework, and frequency of use of tablet devices at home. To construct school-level controls, before construct-

ing averages, we first impute missing values at the individual level using the median response for continuous variables and construct a "missing" category for categorical variables.

3.2 Sample Selection

The Grade 6 LOMCE exams were taken in 1,295 schools in the Community of Madrid, of which 751 lay outside the Capital district and therefore formed the randomized component of the design. Roughly 90% of enrolled students completed the exams, a rate broadly consistent with typical daily absences. We impose three restrictions: non-missing bilingual status, since randomization was stratified on it (Section 2.2); valid scores in all four exams, so that the estimation sample is constant across specifications; and a minimum of 30% of enrolled students and seven students per school. This leaves 719 schools (520 public, 157 charter, and 42 private) and 34,104 students, of whom 21% attend externally evaluated schools—the treatment share in the original experimental sample.

As discussed in Section 5.1, our preferred specifications further restrict attention to the North and South districts, where treatment and control groups are best balanced on observables. This yields the final analysis sample of 408 schools (305 public, 86 charter, and 17 private) and 18,580 students. Appendix Table B.1 reports the number of schools and students surviving each step, separately for each territorial area.

4 Conceptual and Empirical Frameworks

We start by providing a simple model that highlights how heightened moral hazard incentives for schools in the presence of internal evaluation may lead to distortions in school rankings under asymmetric manipulation incentives. Then, we explain how the framework maps into our institutional context and how we empirically evaluate the implications of the model.

4.1 A Model of Strategic Manipulation and Monitoring

Set-up. We consider two types of schools, indexed by $i \in \{P, N\}$, where P denotes public schools and N denotes non-public schools. Each school is characterized by an underlying "true" quality level q_i. However, observed performance (m_i) may differ from the true quality (q_i) due

to endogenous manipulation; that is,

$$m_i = q_i + s_i,$$

where $s_i \geq 0$ denotes manipulation effort. Manipulation in our context is interpreted broadly and does not necessarily require outright falsification of outcomes. Instead, it also encompasses more subtle forms of behavior that can systematically inflate measured performance, such as applying lenient proctoring and grading practices (e.g., providing additional time or undertaking less strict supervision during exams).

Manipulation is costly. Each school type i faces an effort cost of manipulation $C_i(s_i)$ that is increasing in manipulation effort ($C_i'(\cdot) > 0$) and is convex ($C_i''(\cdot) > 0$). These costs reflect both technological and institutional features that discipline manipulation. In addition to the administrative resources required to distort or manipulate outcomes, they also include expected penalties (through sanctions or audits) and reputational costs associated with detection.

Schools derive benefits from manipulation because reported performance determines rankings and, in turn, affects the allocation of resources, reputation, and demand for enrollment. Suppose that the benefits from manipulation are proportional to manipulation effort so that

$$B_i(s_i) = b_i s_i,$$

where $b_i > 0$ captures the marginal benefit from improving measured performance.

Benefits may be heterogeneous. In our particular setting, non-public schools have stronger incentives to manipulate observed performance so that $b_N > b_P$. As we explain above, this is consistent with non-public schools having their funding, enrollment, and survival more tightly linked to measured outcomes, thereby amplifying the returns to improvements in reported performance relative to public schools.

Endogenous manipulation. Each school chooses manipulation effort to maximize net benefits:

$$\max_{s_i \geq 0} \; b_i s_i - C_i(s_i).$$

Under the maintained assumptions on the cost function $C_i(\cdot)$, the first-order condition implicitly defines a unique optimal manipulation level:

$$s_i^* = (C_i')^{-1}(b_i).$$

This characterization yields the standard comparative statics: manipulation increases with the marginal benefits and decreases with the marginal cost. In particular, if costs functions are identical across types, $C_P'(s) = C_N'(s)$, then the assumption $b_N > b_P$ implies that

$$s_N^* > s_P^*.$$

That is, non-public schools, whose payoffs are more tightly linked to reported performance, have stronger incentives to trade-off costs against manipulation and therefore choose higher manipulation levels.

Implications for rankings. Rankings are typically constructed using observed performance m rather than true quality q. To study rankings, we allow quality to vary across schools of the same type: the quality of public schools is distributed according to G_P, and that of non-public schools according to G_N. Since benefits and costs differ only by type, all schools of type i choose the same manipulation level s_i^*, and the measured performance of a school of type i with quality q is $m = q + s_i^*$.[11]

A ranking depends only on the order of measured performance. Two schools of the same type keep their true order, because their scores shift by the same amount. A non-public school with quality q_N ranks above a public school with quality q_P if and only if

$$q_N + \Delta > q_P, \qquad \text{where } \Delta \equiv s_N^* - s_P^*$$

is the manipulation gap. Relative to a ranking by true quality, measured performance therefore behaves as if every non-public school had its quality inflated by Δ. Misranking arises whenever a public school is truly better than a non-public school, but by less than Δ. Since $b_N > b_P$,

[11] If the marginal benefit of manipulation also depended on quality $b_i(q)$, for instance, because the returns to reaching the top of the ranking are larger, manipulation would vary across schools of the same type, and the ranking could also change within type. Implications 3 and 4 below continue to hold as long as monitoring lowers the measured performance of every non-public school by at least as much as that of every public school. If $b_i(q)$ increases with q, manipulation is larger among higher-quality schools, which is consistent with the larger effects of external evaluation in the upper half of the non-public distribution (Section 6). See Appendix D.6.

$\Delta > 0$, and manipulation distorts the ranking in favor of non-public schools.[12]

Two features of the ranking summarize this distortion. First, the probability that a randomly chosen non-public school ranks above a randomly chosen public school,

$$\pi(\Delta) = \Pr(q_N + \Delta > q_P) = \int G_P(q + \Delta)\, dG_N(q),$$

is increasing in Δ. If λ denotes the share of non-public schools, the average percentile rank of non-public schools equals $\lambda/2 + (1 - \lambda)\,\pi(\Delta)$, so it also increases with Δ, while that of public schools decreases. Second, for any top group of the ranking, such as the top 10%, the share of non-public schools in that group increases with Δ. Appendix D derives these results.

The same expression shows where rank changes concentrate. A change in the manipulation gap moves the percentile rank of a non-public school with quality q by approximately $(1 - \lambda)\, g_P(q + \Delta)$ times the change in Δ, where g_P is the density of public-school quality. The same change in manipulation therefore moves a school further in the ranking when many public schools have similar measured performance, as in the middle of the distribution, and moves it little in the tails.

Moral hazard and monitoring. Since reported performance is rewarded, schools operate in an environment where manipulation may be privately optimal even if it does not improve actual learning outcomes. This is a classic moral hazard problem: schools reallocate effort towards activities that raise measured performance, rather than exclusively towards activities that improve true quality. From an external perspective (such as that of parents or policymakers), this implies that rankings may be a misleading signal of school quality and these distortions may have real consequences. For instance, resources may be misallocated across schools if these rankings guide decisions of funding or student demand.

One way policymakers may curb this behavior is by increasing the degree of monitoring, for instance through external supervision of the evaluation process. We incorporate monitoring into the model by letting $f > 0$ denote monitoring intensity, which raises the marginal cost of manipulation. Specifically, the cost function becomes $C_i(s_i, f)$, with

$$\frac{\partial^2 C_i(s_i, f)}{\partial s_i \partial f} > 0,$$

[12] An important benchmark is the case where incentives are symmetric, which happens when cost functions are the same ($C_P(s) = C_N(s)$) and marginal benefits are also equal ($b_P = b_N$). In this case, $s_P^* = s_N^*$, so $\Delta = 0$ and the ranking by measured performance coincides with the ranking by true quality.

implying that higher monitoring intensity makes manipulation more costly at the margin. Within this framework, a shift from internal to external test supervision and grading can be interpreted as an increase in f.

The first order condition in this extension is:

$$b_i = \frac{\partial C_i(s_i, f)}{\partial s_i},$$

where an increase in f shifts up marginal costs, leading to schools reducing manipulation effort in equilibrium. Differentiating the first-order condition gives

$$\frac{\partial s_i^*}{\partial f} = -\frac{\partial^2 C/\partial s\, \partial f}{\partial^2 C/\partial s^2},$$

evaluated at s_i^*. In our context, non-public schools exert more manipulation effort in equilibrium ($s_N^* > s_P^*$). With a common cost function, monitoring therefore reduces manipulation more among non-public schools if it raises marginal costs more strongly at higher levels of manipulation, relative to the curvature of costs (i.e., the ratio $(\partial^2 C/\partial s \partial f)/(\partial^2 C/\partial s^2)$ is increasing in s, as when costs are quadratic in s and $\partial^2 C/\partial s \partial f$ is increasing in s). In that case,

$$\left|\frac{\partial s_N^*}{\partial f}\right| > \left|\frac{\partial s_P^*}{\partial f}\right|.$$

Hence, a tightening of monitoring leads to a larger reduction in reported performance for non-public schools which leads to improvement in the alignment between observed outcomes and true quality.

Since manipulation falls faster for non-public schools, monitoring narrows the manipulation gap:

$$\frac{\partial \Delta}{\partial f} = \frac{\partial s_N^*}{\partial f} - \frac{\partial s_P^*}{\partial f} < 0.$$

External evaluation therefore reduces the distortion of the ranking in favor of non-public schools.

4.2 Testable implications

Based on the stylized model described above, there are four implications that we are able to test in the data. The first two concern average scores: we should observe a gap in average scores between external and internal evaluation, and this gap should depend on manipulation incentives

and, in turn, on ownership structure. The last two concern rankings: external evaluation should lower the position of non-public schools in the ranking and reduce their presence at its top.

Implication 1: Measured performance is on average lower under external evaluation. Since external evaluation increases monitoring intensity, it effectively raises the marginal cost of manipulation. This implies that optimal manipulation weakly decreases $s^*(f_{ext}) \leq s^*(f_{int})$. Holding true performance fixed, this implies

$$E[m \mid \text{External} = 1] \leq E[m \mid \text{External} = 0]. \tag{1}$$

Implication 2: Larger score differences are observed among those whose incentives are stronger. Under weaker monitoring (internal evaluation), non-public schools choose higher manipulation $s^*_N > s^*_P$. Under stronger monitoring (external evaluation), the reduction in manipulation is larger where the baseline is larger. This yields:

$$E[m \mid \text{Ext} = 1, N] - E[m \mid \text{Ext} = 0, N] < E[m \mid \text{Ext} = 1, P] - E[m \mid \text{Ext} = 0, P]. \tag{2}$$

Implication 3: Non-public schools rank lower under external evaluation. Since external evaluation narrows the manipulation gap, $\Delta(f_{ext}) < \Delta(f_{int})$, the probability that a non-public school ranks above a public school is lower under external evaluation, $\pi(\Delta(f_{ext})) < \pi(\Delta(f_{int}))$. Equivalently, the average percentile rank R of non-public schools is lower:

$$E[R \mid \text{Ext} = 1, N] < E[R \mid \text{Ext} = 0, N]. \tag{3}$$

Implication 4: Non-public schools are less represented at the top of the ranking under external evaluation. For any top group of the ranking, the share of non-public schools increases with Δ. Hence, under external evaluation,

$$\Pr(N \mid \text{top } x\%, \text{Ext} = 1) < \Pr(N \mid \text{top } x\%, \text{Ext} = 0) \quad \text{for any } x. \tag{4}$$

Since true quality is not observed, we cannot measure manipulation directly. Instead, we test these implications empirically by exploiting the random assignment of schools to evaluation regimes: Section 5.2 estimates the average effect of external evaluation on measured performance (Implication 1) and how this effect differs by school ownership (Implication 2), using the em-

pirical framework described next. Section 6.3 tests Implications 3 and 4 using rank regressions and counterfactual rankings to examine whether external evaluation lowers the score ranking of non-public schools and reduces their representation at the top of the score distribution.

4.3 Empirical Framework

Design-based identification and inference. We exploit random assignment of all primary schools outside the Capital district to internal versus external evaluation of Grade 6 LOMCE exams, with stratification by district, school ownership, bilingual-program status, and school size. Because treatment is randomly assigned at the school level, differences in outcomes between treated and control schools identify causal effects of external evaluation. Our baseline specifications include strata fixed effects to account for the stratification procedure outlined. Unless otherwise specified, we follow Abadie et al. (2023) and conduct inference at the level of treatment assignment. Since our main regressions are estimated at the school level, which is also the level of treatment assignment, we use heteroskedasticity-robust standard errors, which, with one observation per school, are equivalent to clustering at the school level. In cases where the unit of observation is more disaggregated, we cluster standard errors at the school level.

Baseline specification: Average treatment effects. Motivated by Equation (1) in our conceptual framework, we begin by estimating the average effect of external evaluation on test scores using the following specification:

$$Y_i = \theta_0 + \theta_1 \text{External}_i + X_i\gamma + Z_i\delta + \varepsilon_i, \tag{5}$$

where Y_i denotes the average standardized score in a given subject for school i, and External_i is an indicator equal to one if the school is assigned to external evaluation. The vector X_i contains average student characteristics at the school level, which are included to improve precision. The vector Z_i captures the stratification variables used in the experimental design: fixed effects for the strata defined by district, bilingual status, and ownership (public or non-public), and their interactions with school size, which we measure by the number of students in grade 6.

Our main explanatory variable is an indicator equal to one for external evaluation. Because external evaluation was the status-quo regime, this coding allows us to interpret the coefficient θ_1 as the effect of maintaining the status quo rather than switching to internal evaluation. This specification corresponds directly to Implication 1, which predicts that external monitoring

reduces measured performance by limiting manipulation ($\theta_1 < 0$).

Heterogeneity by incentives. Motivated by Equation (2) in our conceptual framework, to test whether the effect of monitoring varies with incentives, we allow for heterogeneous treatment effects by school ownership:

$$Y_i = \beta_0 + \beta_1 \text{External}_i + \beta_2 \text{Non-public}_i + \beta_3 \text{External}_i \times \text{Non-public}_i + X_i\gamma + Z_i\delta + u_i, \quad (6)$$

where Non-public_i is an indicator for charter and private schools (with public schools as the omitted reference category). Because ownership is one of the randomization strata, β_2 is absorbed by the strata fixed effects in Z_i, and we do not report it.

We are interested in the coefficient of the interaction β_3, which captures the differential impact of external evaluation on non-public schools. Under the conceptual framework, we expect $\beta_3 < 0$, reflecting stronger manipulation incentives in non-public schools and therefore a larger reduction in scores under external monitoring (Implication 2).

Rankings. Motivated by Equations (3) and (4), we also estimate Equation (6) with the school's percentile rank, or an indicator for ranking in the top of the distribution, as the outcome. Under the conceptual framework, we again expect $\beta_3 < 0$. We complement these rank regressions with counterfactual rankings under external evaluation, and describe these analyses in Section 6.3.

Additional analyses. To support and interpret the model-implied effects, we complement the baseline analysis with other empirical exercises. In particular, we explore how external evaluation affects the distribution of scores and participation in the exam, and we assess whether selective participation can account for our results. We discuss their implementation and interpretation where relevant throughout the paper.

5 Main Results and Robustness

5.1 Balance Checks

We begin by assessing whether the randomization achieved balance in observable characteristics between treatment and control groups. Specifically, we estimate the following equation:

$$X_i = \alpha_0 + \tau \times \text{External}_i + Z_i\delta + \varsigma_i, \tag{7}$$

where X are key observable characteristics for which we will check balance and the rest mimics Equation (5). ς is an error term.

In Table B.2, a systematic evaluation of the 719 schools outside of the Capital region reveals that, while many characteristics are balanced, schools that were externally evaluated in 2017 had performed better in 2016 than their internally evaluated counterparts. We further show that this is fully driven by schools in the East and West regions, and that the balance is completely maintained in the North and South regions (Table B.3). In light of this, we conservatively focus on the subsample of Northern and Southern schools — a similar approach is followed in Montalbán-Castilla and Sevilla (2019). However, we have verified that all results hold when also including the East and West regions.[13]

The balance checks for the Northern and Southern schools are reported in Table 2. Panel A shows small and statistically non-significant differences across key drivers of academic performance, including (i) predetermined characteristics (such as gender and country of origin) and (ii) endogenously determined family-level investments (help with homework, use of tablets at home) and child behavior (frequency of doing homework and absences). School ownership, the main focus of the paper, is balanced by design, since it is one of the randomization strata (column 2). We only find a statistical difference in repetition, with students assigned to external grading being less likely to have experienced grade retention. Reassuringly, the economic magnitude of the difference is small. Moreover, given the direction of the difference, if anything, this would go against our main findings as it suggests that externally evaluated students are better performers on average. It is also worth noting that, given the number of covariates examined, an isolated statistically significant difference may arise by chance even under successful randomization. In any case, this difference will be accounted for through regression adjustment. Importantly, we do not find any difference in the schools' test scores across treatment and control in 2016, the

[13] For brevity, we document some checks in the Appendix and any further analysis is available upon request.

Table 2: Balance Check

(a) Individual Characteristics

	(1) Female	(2) Non-public	(3) Non-Span.	(4) Parent help	(5) HW freq.	(6) Tablet home	(7) Absences	(8) Repeater
External	-0.017	–	0.001	-0.008	0.002	-0.005	-0.006	-0.020*
	(0.011)	–	(0.013)	(0.019)	(0.006)	(0.012)	(0.004)	(0.011)
Observations	406	408	401	401	401	401	377	401
R-squared	0.011	1.000	0.036	0.012	0.041	0.032	0.041	0.188

(b) Missing Individual Characteristics

	(1) Female miss.	(2) Non-public miss.	(3) Non-Span. miss.	(4) Parent miss.	(5) HW freq. miss.	(6) Tablet miss.	(7) Absences miss.	(8) Repeater miss.
External	0.010	–	-0.008	-0.008	-0.011	-0.009	-0.027	-0.009
	(0.013)	–	(0.032)	(0.032)	(0.032)	(0.032)	(0.041)	(0.032)
Observations	408		408	408	408	408	408	408
R-squared	0.045		0.019	0.018	0.019	0.019	0.021	0.018

(c) 2016 School Test Scores

	(1) Composite	(2) Spanish	(3) English	(4) Mathematics	(5) Technology
External	1.502	3.506	-1.101	1.791	1.810
	(4.279)	(5.080)	(3.962)	(5.152)	(4.959)
Observations	405	405	405	405	405
R-squared	0.186	0.101	0.491	0.115	0.094

Notes: Balance tests following Equation (7). "External" is an indicator taking the value of 1 if the school was externally evaluated and 0 otherwise. Outcome throughout the table are averaged at the school level. Outcomes in Panel B are the proportion of individuals within the school for which the variable was missing. "Non-public miss" is never missing since this information being available at the school level is a sample selection criterion. Panel C has as outcomes the average school score in the LOMCE exam for the 2016 academic year (the previous one to the RCT we evaluate). Robust standard errors clustered at school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

year before our RCT.

As discussed, we will impute the value of the controls that are missing at the individual level. In Panel B, we therefore verify that there are no differences across treatment and control in the availability of the key variables from Panel A. This also serves to further strengthen the underlying comparability of these two groups.[14]

[14] A short discussion on compliance with the treatment (i.e. potential selection into/out of treatment on the part of the schools and/or the students) is relevant. The possibility of worse schools opting out from treatment does not apply in this context since schools that were assigned to treatment was forced to comply with it by the Community of Madrid. On the other hand, students' self-selection may take place since the socio-demographic characteristics of the students are provided by their parents through answering a voluntary online survey. However, assuming that the choice of filling out the survey is based on observable characteristics, it is not clear why it should be the case that parents in treated and untreated schools should have different response rates. Table B.4 shows that the ratio of responses to the survey between treated and untreated schools is not different by taking the ratio of the number of students with survey information over the total number of students enrolled.

Table 3: Effects of External Grading on Academic Scores, Grade 6 (2017)

	No Controls	**With Controls**
Panel A		
External Grading	-9.463**	-11.465***
	(4.380)	(3.962)
Observations	408	408
R-squared	0.265	0.393
Mean (sd) Dep. Var. Control	489.140 (44.293)	489.140 (44.293)
Panel B		
External	-0.985	-1.968
	(4.430)	(4.028)
External × Charter/Private	-32.722***	-36.492***
	(11.073)	(9.513)
Observations	408	408
R-squared	0.282	0.413

Notes: Regressions follow Equation (5) in Panel A and Equation (6) in Panel B. The left part does not include additional controls beyond the stratification variables. The right part includes as controls the school-level averages of the variables in Table 2's Panel A. Panel B pools non-public schools, i.e., charter and private, into a single category ("Charter/Private"). The omitted category are public schools. Robust standard errors clustered at school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.2 Effects of External Evaluation and Heterogeneity by Incentives

Panel A of Table 3 reports the estimates, following Equation (5), of the average effect of assignment to external proctoring and evaluation on LOMCE scores. Across subjects (Panel A of Table 4), externally evaluated schools perform systematically worse than internally evaluated ones. The effects are particularly large and significant for the two quantitative subjects (mathematics and technology) in specifications without controls, and they also become significant for Spanish once controls are included (reflecting both an increase in the point estimate and in

Table 4: Effects of External Grading on Academic Scores by Subject, Grade 6 (2017)

	No Controls				**With Controls**			
	Spanish	English	Math	Tech.	Spanish	English	Math	Tech.
Panel A								
External Grading	-8.581	-3.242	-14.730***	-11.298**	-10.493**	-4.193	-18.233***	-12.941***
	(5.451)	(5.412)	(5.573)	(4.760)	(4.858)	(5.129)	(5.585)	(4.629)
Observations	408				408			
R-squared	0.185	0.470	0.175	0.113	0.348	0.557	0.262	0.189
Mean (sd) Dep. Var. Control	490.862 (44.773)	482.624 (58.847)	490.137 (58.419)	492.937 (51.412)	490.862 (44.773)	482.624 (58.847)	490.137 (58.419)	492.937 (51.412)
Panel B								
External	-3.273	4.622	-1.304	-3.987	-4.459	5.224	-3.929	-4.708
	(5.835)	(6.010)	(5.496)	(5.008)	(5.322)	(5.794)	(5.508)	(4.926)
External × Charter/Private	-20.488	-30.355**	-51.829***	-28.218**	-23.186*	-36.186***	-54.964***	-31.634***
	(13.827)	(12.572)	(13.599)	(11.940)	(11.926)	(11.475)	(12.583)	(11.095)
Observations	408				408			
R-squared	0.191	0.478	0.201	0.123	0.355	0.567	0.290	0.201

Notes: Regressions follow Equation (5) in Panel A and Equation (6) in Panel B. The left part does not include additional controls beyond the stratification variables. The right part includes as controls the school-level averages of the variables in Table 2's Panel A. Panel B pools non-public schools, i.e., charter and private, into a single category ("Charter/Private"). The omitted category are public schools. Robust standard errors clustered at school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

precision).

These results are consistent with the presence of score inflation under internal evaluation, whereby weaker monitoring allows schools to enhance measured performance without necessarily improving underlying learning. The fact that the negative effects arise despite random assignment and remain robust to the inclusion of controls suggests that they reflect a causal impact of the evaluation regime rather than differences in school or student composition. Overall, these baseline estimates provide initial evidence that external monitoring reduces measured performance, consistent with the removal of manipulation highlighted in the conceptual framework.

Heterogeneity of effects by school ownership. Panel B of Table 3 augments the baseline specification by interacting external evaluation with an indicator for non-public (pooling private and charter) schools. There is clear evidence that the average negative effect of external evaluation is driven by non-public schools: once interactions are included, the coefficient on external evaluation alone becomes small and statistically insignificant, while the interaction terms are consistently negative and often sizeable, indicating that non-public schools experience significantly larger score reductions under external monitoring. This pattern aligns closely with the model's prediction that schools facing stronger incentives to appear high-performing engage more in manipulation when monitoring is weak, and therefore exhibit larger declines when monitoring is tightened. Taken together, these results suggest that heterogeneity in incentives plays a central role in shaping the response to external evaluation and is key to understanding the aggregate effects documented in Panel A.

Heterogeneity by bilingual status. An incentive-based interpretation also predicts that the ownership gradient should be stronger in contexts where performance in a certain subject may carry particular reputational value. We examine this prediction using scores in the English section between schools in the English bilingual program. This program, which requires schools to use English as the language of instruction for at least 30% of teaching time (and is only available for public and charter schools), has become a salient margin of differentiation in the education market in Madrid. For schools operating under this regime, performance in English is likely to be particularly visible and valuable for parents, thereby strengthening the incentives to boost recorded achievement in this subject. Consistent with this mechanism, one would expect the effects of external evaluation in English to be especially pronounced among bilingual schools,

and in particular among non-public bilingual schools, which rely more heavily on comparative performance for student demand and funding.

In Table B.5, we replicate Panel B of Table 3 for English, separately for bilingual and non-bilingual schools. The differential effect of external evaluation on non-public schools is substantially larger among bilingual schools (column 1) than among non-bilingual schools (column 2), both for the overall English score and for its written part. The pattern is the same when we drop private schools, which are not officially part of the program (columns 3 and 4). The larger ownership gradient among bilingual schools, particularly in written English, provides additional support for the interpretation that reputational incentives contribute to score manipulation under internal evaluation.

5.3 Robustness and Additional Analyses

The RCT design provides a compelling basis for causal inference and the credibility of the randomization is further supported by the balance in observables documented in Table 2. In this section, we assess the robustness and sensitivity of our main findings to alternative specifications and samples, and present additional analyses that complement the baseline results.

Distinguishing charter and private schools. Our baseline specification pools charter and private schools into a single non-public category. We augment our preferred specification by including separate indicators for charter and private schools and interacting each of them with external evaluation. Appendix Table B.6 shows that the negative differential effect of external evaluation is present for both charter and private schools. The estimated interaction is consistently negative for charter schools and is even larger in magnitude for private schools, although the latter estimates are less precise across some subjects. We retain the pooled specification as our baseline because the sample contains only 17 private schools, limiting power and making those estimates sensitive to individual observations. These results, however, indicate that our main findings are not driven exclusively by either type of non-public school and support the broader interpretation that schools facing stronger competitive and reputational incentives experience larger reductions in measured performance under external evaluation.

Randomization inference. To assess whether our results could arise by chance under the randomization design, we implement randomization inference (Imbens and Wooldridge, 2009).

This approach compares our estimated effects to the distribution obtained under repeated random assignments of treatment, providing a design-based benchmark for statistical significance that does not rely on large-sample approximations. For this exercise, we use 1,000 permutations. Figure A.3 shows that our results are consistent with those in Tables 3 and 4. The estimated interactions for the composite score, English, mathematics, and technology fall outside the 90% two-tailed intervals implied by the permutation distributions, and the estimate for Spanish lies at the endpoint of the corresponding interval.

Controlling for past performance at the school level. To further strengthen the comparability between the treatment and control groups, we control for lagged performance by including the scores obtained by the previous Grade 6 cohort in each school on the standardized evaluations administered during the preceding academic year. These scores provide a summary measure of underlying school quality and student composition, capturing persistent factors such as demographics, school inputs, and unobserved institutional characteristics that may not be fully accounted for by the observed controls. Table B.7 shows that our main results are largely unchanged.

Controlling for socioeconomic status. Our controls are school averages of individual student characteristics. The education authority also reports, for each school, an index of socioeconomic and cultural status (ISEC), built from parental education and occupation, household resources, and the number of books at home. Table B.8 shows that controlling for this index leaves the ownership gradient essentially unchanged: the interaction for the composite score is -29.8 without and -33.3 with our other controls, against -32.7 and -36.5 in Table 3. We do not use the index in our main specification because its components come from the voluntary family questionnaires, which only about half of families completed (Section 3). For many schools it therefore reflects a selected subset of families, and it is missing for one school.

Number of correct answers. The official scores are scaled and standardized by the education authority. As an alternative measure of performance that does not depend on this scaling, Table B.9 uses the school-average number of correct answers in each subject. The interaction between external evaluation and non-public ownership is negative in every subject and, with controls, statistically significant in all of them, as well as for a composite that averages the four standardized counts.

Analysis at the student level. Our baseline estimates use school-level average outcomes, matching the unit of treatment assignment and assigning equal weight to each school. As a robustness check, we re-estimate the main specifications at the student level. This alternative specification weights schools according to the number of students observed and allows us to verify that the results are not driven by the aggregation or equal weighting of schools in the baseline analysis.[15] Table B.10 shows that the results are consistent across both levels of analysis. In the student-level regressions, we cluster standard errors at the school level to account for the common treatment assignment and dependence among students within schools.

Analysis by district. To verify that our findings are not driven by a particular geographic area, we estimate the main specification separately for each district in the baseline sample. This exercise allows us to assess the consistency of the effects across districts with different socioeconomic compositions. Table B.11 shows broadly similar patterns across regions (the effects for North are less precisely estimated as it is a smaller district).

Including all districts (except Capital). Our preferred specification focuses on districts where balance is strongest, minimizing concerns about residual differences across treatment groups. To assess the sensitivity of our results to this choice, we replicate the analysis using all districts included in the RCT (i.e., adding East and West, since Capital was not part of the randomization). Table B.12 shows that the main results hold. This indicates that our findings are not driven by the specific sample used in the baseline specification. These results also speak to the plausible external validity of our baseline specification, which focuses on two regions, by showing that our findings extend to the broader Community of Madrid.

Complementary evidence from grade 10 tests. A similar random assignment to internal and external evaluation was implemented in the 2017 Grade 10 LOMCE exams, which provides an opportunity to examine whether the main prediction of our framework extends to a different level of education. However, the administrative score records are substantially less complete than in Grade 6, with missing and partial records differing across school ownership and evaluation regimes.[16] These limitations prevent us from establishing the comparability and balance

[15] When regressors vary only at the school level, weighted least squares on school-level means, using the number of observed students as weights, reproduces the corresponding student-level OLS estimates. This equivalence need not hold when the specifications include student-level regressors or use different weighting schemes.

[16] In Table C.1, we show that partial records are disproportionately concentrated among externally evaluated public-school students and are unusually low-scoring, making estimates sensitive to their treatment.

that underpin our Grade 6 analysis. We therefore treat this exercise as complementary evidence rather than as part of our main analysis, and provide full details of the analysis in Appendix C.

Consistent with our main results, we find the same ownership gradient: external evaluation is associated with a larger decline in scores among non-public schools. The gradient is most robust in mathematics, where it does not depend on how incomplete score records are treated; for the composite score, it is present only when students with scores in some but not all subjects are excluded, as in our Grade 6 sample rules. These findings provide additional, albeit suggestive, support for the interpretation that heterogeneous incentives shape responses to monitoring.

5.4 Discussion: School Manipulation versus Student and Family Responses

Our experimental design identifies the causal effect of external evaluation on measured performance, but it does not isolate the precise mechanisms generating this effect. Internal evaluation may create scope for several forms of strategic behavior by schools. These need not involve the direct alteration of students' answers (Jacob and Levitt, 2003). They may include more lenient supervision or grading, additional time, assistance during the exam, or changes in which students participate (Figlio and Getzler, 2006; Angrist et al., 2017; Dee et al., 2019; Lucifora and Tonello, 2015). We do not attempt to separate these channels, although the open-ended writing sections of the Spanish and English exams, where graders have the most discretion, are suggestive: external evaluation lowers writing scores in both types of schools, and by 7.0 percentage points of the maximum score more in non-public schools (Appendix Table B.13), which is consistent with more lenient grading of open-ended answers under internal evaluation as one of the active mechanisms. We view all of the different school actions as responses of the evaluated school to weaker monitoring, which is the behavior our framework describes and the one that distorts comparisons across schools. The relevant concern is instead that external evaluation may change the behavior of students and their families, which would not be manipulation by schools. We consider whether external proctors affect students' performance or whether students and their families adjust exam participation. We argue below that neither channel is likely to account for the observed ownership gradient.

Behavioral responses of students. Being assessed by unfamiliar proctors may induce stress, anxiety, or distraction, thereby reducing performance (Bertoni et al., 2013; Cai et al., 2019), whereas students may instead exert more effort if they perceive the exam as more con-

sequential (Schlosser et al., 2019). For such responses to explain our results, they must account not merely for lower average scores under external evaluation, but also for the pronounced gradient by school ownership. This would require substantially stronger adverse responses among students in non-public schools, despite common exam content, timing, and administrative procedures across schools. While we regard these differences as unlikely, in our main results, we control for average student characteristics to account for observed differences in student composition that might otherwise explain differential responses to external proctors across ownership types. Moreover, Table 3 shows a small and statistically insignificant effect of external evaluation among public schools, offering little evidence of a net performance response in that sector. Overall, we therefore find little evidence that the behavioral channel is at work.

Selective participation by students and families. Selective participation is a more salient concern in our setting. The LOMCE examinations were controversial, and contemporary reports document explicit calls by education unions, families, and student organizations in Madrid for students not to sit them.[17] These calls make it plausible that participation responded through actions by schools or teachers, decisions by students and their families, or both. The distinction matters for interpretation. Schools under accountability pressure may strategically shape the pool of test takers (Figlio and Getzler, 2006; Cullen and Reback, 2006; Lemke et al., 2006), and such actions would be part of the response to monitoring that we aim to capture. By contrast, independent participation decisions by students and families could generate score differences without school manipulation.

Table B.14 reports the effects of external evaluation on exam non-participation.[18] External evaluation increases non-participation in non-public schools relative to public schools. Whether this response can explain the ownership gradient in scores depends on which students do not participate. Studies of test exclusion find that non-participants tend to be low achievers, whether through exemptions, reclassification as disabled, special-education placement, grade retention, or disciplinary suspensions timed around the test (Figlio and Getzler, 2006; Cullen and Reback, 2006; Jacob, 2005; Figlio, 2006). If non-participants in our setting are likewise weaker than those who sit the exam, their absence would raise the measured scores of externally evaluated non-public schools, attenuating rather than explaining our main result.

[17] See, for instance, https://www.madridiario.es/444357/comunidad-educativa-revalidas-lomce-boicot or https://www.elmundo.es/sociedad/2016/05/06/572c8390468aeb29348b4628.html.

[18] The average absence rate in our estimation sample is 6.3%, computed as the ratio of students missing the test to those enrolled.

We assess the sensitivity of our findings to alternative patterns of selection using Lee (2009) bounds, reported in Table B.15. Participation differences between evaluation regimes are small, about two percentage points within each ownership type, and their sign varies across randomization strata, so we compute the bounds within strata. Under the maintained selection assumptions, these bounds identify a range of treatment effects among students who would sit the exam under either evaluation regime. The bounds for non-public schools remain negative and economically meaningful. More importantly, the bounds on the difference in treatment effects between non-public and public schools are entirely negative: between -34 and -18 points for the composite score, and below zero in every subject. Thus, a negative ownership differential remains even when allowing for selective participation.

These findings suggest that participation is a relevant response to external evaluation but does not fully explain the ownership gradient in scores. If weaker students disproportionately select out, the participation response works against our results. Moreover, when we allow for alternative patterns of selection, the bounds leave a substantial negative ownership differential among always-participants. If schools drive these absences, participation is itself a margin of strategic behavior. If students and families drive them independently, the remaining differential is consistent with school manipulation through other channels, including invigilation and grading, although it does not by itself establish those mechanisms.

6 Distributional Effects and School Rankings

Our conceptual framework highlighted that when measured performance reflects both true achievement and manipulation, rankings based on observed scores misrepresent relative quality. While the average effects we have estimated show that internal evaluation raises the scores of non-public schools, they do not tell us where in the distribution this happens, which is what rankings depend on. In this section, we trace how the evaluation regime shapes the distribution of school-level average scores and, in turn, school rankings.

6.1 Evidence from quantile regressions

We estimate quantile regressions of the school-level average score on the external-evaluation indicator, the non-public indicator, their interactions, and the same controls as the main average results at the 10th, 25th, 50th, 75th, and 90th percentiles. Table 5 reports the estimates

for the composite score, and Appendix Table B.16 for each subject. The coefficient of the external indicator is the effect of external evaluation on a given conditional quantile of the score distribution of public schools, and the interaction is the additional effect for non-public schools.

Table 5: Quantile Treatment Effects

	(1) q10	(2) q25	(3) q50	(4) q75	(5) q90
External	11.147**	3.036	-1.318	-4.988	-13.091
	(5.623)	(4.606)	(5.088)	(7.580)	(9.915)
Ext. × Priv./Conc.	-27.927**	-15.781	-28.904***	-38.940***	-33.013*
	(14.148)	(13.178)	(10.498)	(13.736)	(18.292)
Pseudo R-squared	0.326	0.265	0.267	0.271	0.297

Notes: Quantile regression estimates at indicated percentiles. All regressions include the same controls as in Table 3. All panels are estimated on 408 observations. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

For non-public schools, the effect of external evaluation is largest in the upper half of the distribution. The interaction is negative at every quantile and largest at the 75th and 90th percentiles. Adding the coefficient on External, external evaluation lowers the composite score of non-public schools by 17 points at the 10th percentile, 30 points at the median, and 44 and 46 points at the 75th and 90th percentiles. For public schools, the effects are small throughout the distribution, apart from a modest increase at the 10th percentile. Internal evaluation therefore raises non-public scores most in the upper half of the distribution, where the top positions of a ranking are decided.

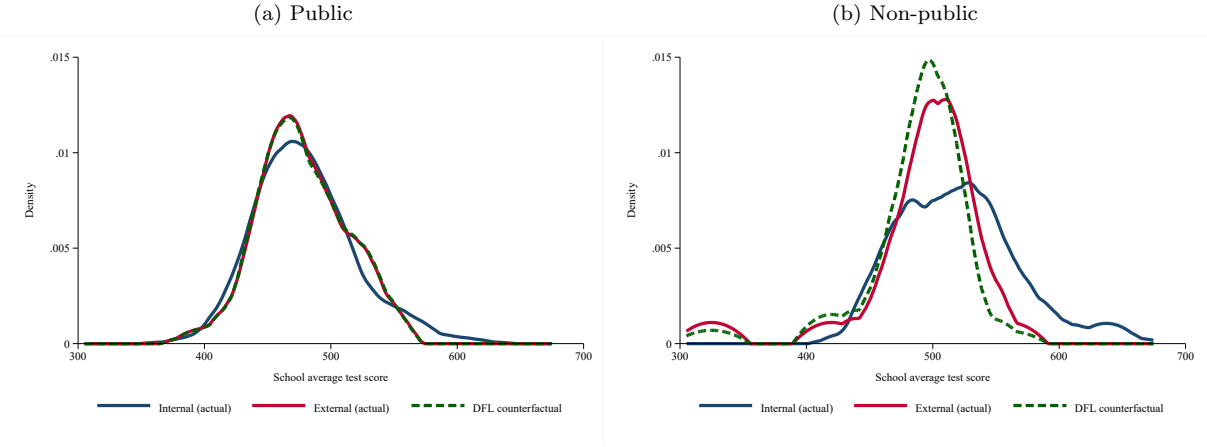
6.2 Counterfactual score distributions

The quantile regressions above show how far specific points in the distribution move, but a ranking depends on how many schools sit in each score range and on where the schools would otherwise lie in alternative evaluation regimes. Moreover, the previous exercise estimate *conditional* quantiles, whereas rankings compare schools in the unconditional distribution, and the two need not coincide (Firpo et al., 2009). To further quantify the implications of evaluation regimes for school rankings, we construct counterfactual score distributions using the reweighting approach of DiNardo et al. (1996) (hereafter referred to as DFL) separately for public and non-public schools.

The DFL exercise reweights externally evaluated schools to match internally evaluated schools of the same ownership type on the variables governing assignment. The resulting

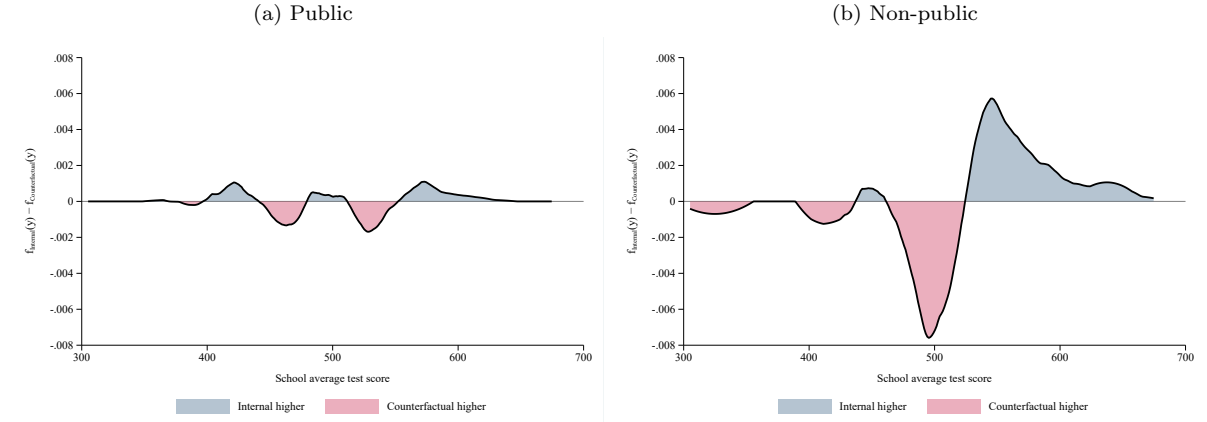
weighted score distribution estimates what internally evaluated schools would have scored under external evaluation. Because evaluation regimes were randomly assigned, reweighting adjusts for finite-sample imbalances rather than corrects for selection, a consideration particularly relevant for non-public schools, since only 22 of which were externally evaluated. We describe this methodology in more detail in Appendix E.

Figure 1: Actual and Counterfactual Distributions of School Scores, Composite Score



Notes: Kernel densities of the school-level average composite score. Internal (actual) and External (actual) are the observed densities under each evaluation regime. The DFL counterfactual is the density of externally evaluated schools reweighted to the observable characteristics of internally evaluated schools of the same ownership type, i.e., the distribution internally evaluated schools would display under external grading. Weights and densities are estimated separately by ownership type, as described in Appendix E. The sample includes 305 public schools (247 internally and 58 externally evaluated) and 103 non-public schools (81 and 22).

Figure 2: Difference between Actual and Counterfactual Densities, Composite Score



Notes: Each panel plots $\Delta(y) = f_{\text{int}}(y) - f_{\text{cf}}(y)$, the actual density under internal evaluation minus the DFL counterfactual density in Figure 1. Blue areas mark scores at which internal evaluation places more schools than the counterfactual, and red areas scores at which it places fewer. Both panels share the same vertical scale.

In Figure 1, we plot, for the composite score, the actual densities under internal and external evaluation together with the counterfactual. For public schools, the three densities are similar (panel (a)). For non-public schools, the density under internal evaluation differs markedly from

both the density under external evaluation and the counterfactual.[19] Under external grading, internally evaluated non-public schools would be concentrated around 500 points, whereas under internal grading they spread from about 460 to 560 points and have a thick right tail (panel (b)).

In Figure 2, we plot the difference between the actual density under internal evaluation and the counterfactual, $\Delta(y) = f_{\text{int}}(y) - f_{\text{cf}}(y)$. $\Delta(y)$ is positive at scores where internal evaluation places more schools than external evaluation would, and negative where it places fewer. Since both densities integrate to one, the positive and negative areas are equal, so $\Delta(y)$ shows from which score ranges internal evaluation removes mass and to which ranges it adds it. Internal evaluation shifts the non-public score distribution from the middle of the distribution toward the upper half. Compared with the counterfactual distribution under external evaluation, the observed distribution under internal evaluation has fewer schools scoring between roughly 460 and 525 points and more scoring above 525, with a tail extending beyond 650 points (panel (b)). For public schools, the difference is small, at most about a fifth of the non-public one, and has no systematic sign (panel (a)). By subject, the same reallocation from the middle to the upper half of the non-public distribution appears in all four subjects; in Spanish, the mass is drawn from the lower part of the distribution as well as from the middle (Appendix Figures A.4 and A.5).

The counterfactual sharpens the quantile evidence in two ways. First, it shows that the larger interactions at the upper quantiles reflect a displacement of mass from the middle of the non-public distribution into its upper half. Second, it locates the region from which this mass is drawn, around and just below the median, which is where schools are most densely packed.

6.3 Implications for school rankings

Differences in score inflation across ownership types may distort school rankings, affecting both schools' relative positions and which schools place at the top. The preceding analyses document changes in score distributions but do not quantify these ranking effects. We examine effects on school rankings through three exercises, testing Implications 3 and 4 of our conceptual

[19] Among non-public schools, the counterfactual also differs somewhat from the actual density under external evaluation. Both are densities of the same 22 externally evaluated non-public schools, so the difference reflects the reweighting rather than the evaluation regime. With so few schools, their design variables differ by chance from those of the 81 internally evaluated non-public schools: for instance, one randomization stratum contains 27% of the externally evaluated non-public schools but only 14% of the internally evaluated ones. The weights correct for these differences, which lowers the mean of the counterfactual slightly, from 493 to 488 points, and makes it more concentrated, with a standard deviation of 40 rather than 48 points. Among public schools, with 58 externally evaluated schools, the design variables are close to balanced and the two densities nearly coincide.

framework (Section 4.2). First, we rank all 408 schools by their average score and estimate how external evaluation affects a school's percentile rank and probability of appearing in the top quartile or decile. Second, we use the DFL counterfactual distributions to assess how universal external evaluation would change the ownership composition at the top of the ranking. Third, we construct school-level counterfactual scores and re-rank schools to examine individual rank changes and identify which schools would enter or leave the top of the ranking.

Effect of external evaluation on a school's rank. We first estimate Equation (6) with controls, replacing the score with the school's percentile rank or with indicators for ranking in the top quartile or the top decile. Table 6 reports the results for the composite score. External evaluation lowers the rank of non-public schools by 15.6 percentile points, from an average of 69.8 among internally evaluated non-public schools. It also lowers their probability of ranking in the top quartile by 34 percentage points and in the top decile by 28 points. Because a school's rank depends on the scores of all other schools, these estimates measure the effect of external evaluation on the rank of a school when most other schools are internally evaluated. They do not tell us how the ranking would change if all schools were evaluated externally.

Table 6: Effect of External Evaluation on School Rank, Composite Score

	Percentile rank (1)	Top quartile (2)	Top decile (3)
External	−0.30	1.86	−8.04
	(3.07)	(5.29)	(2.13)
External × Priv./Conc.	−15.28	−35.85	−20.31
	(5.82)	(10.82)	(6.97)
	[0.021]	[0.001]	[0.045]
External + External × Priv./Conc.	−15.57	−34.00	−28.35
	(4.90)	(9.54)	(6.77)
	[0.008]	[<0.001]	[0.003]
Mean, internally evaluated Priv./Conc.	69.8	53.1	25.9
Observations	408	408	408

Notes: School-level regressions following Equation (6) with the same controls as in Table 3. The outcome is the school's percentile rank (0–100) in the pooled ranking of the 408 schools by composite score in column (1), and an indicator, in percentage points, for ranking in the top 25% or the top 10% in columns (2) and (3). Standard errors clustered at the school level in parentheses. Randomization-inference p-values in brackets.

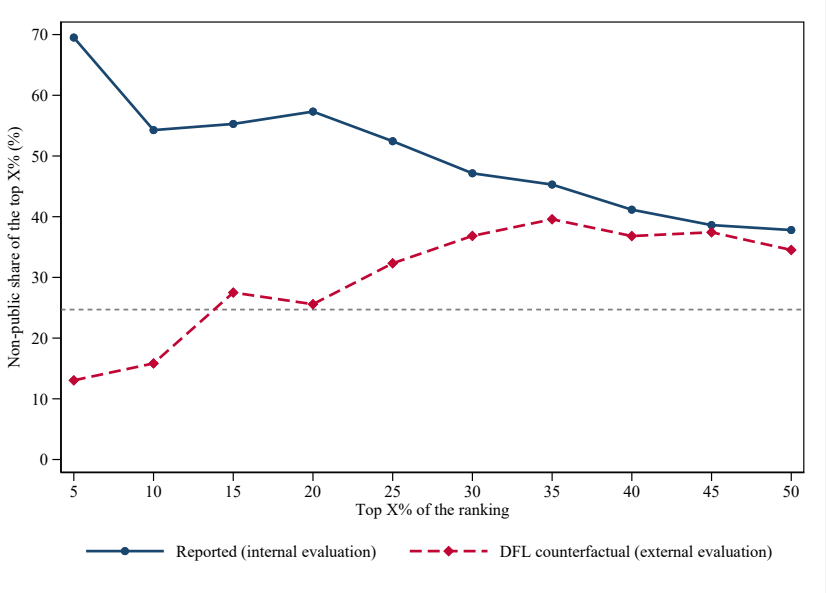
Counterfactual ranking. Rankings depend only on the pooled distribution of scores, so the counterfactual of Section 6.2 also yields the ranking that internally evaluated schools would display under external grading. The reported ranking ranks the 328 internally evaluated schools

by their scores. The counterfactual ranking ranks the 80 externally evaluated schools, each weighted by its DFL weight, its estimated odds of internal evaluation $\hat{\Psi}_s = (1 - \hat{p}(z_s))/\hat{p}(z_s)$ (Appendix E). We rescale these weights within each ownership type so that they sum to the number of internally evaluated schools of that type, 247 public and 81 non-public. The counterfactual ranking therefore describes a population with the same size and ownership composition as the internally evaluated schools, in which the scores of each ownership type follow its counterfactual distribution. In both rankings, the percentile rank of school s is the weighted share of schools with lower scores,

$$R_s = 100 \times \frac{\sum_{r:\, y_r < y_s} w_r + w_s/2}{\sum_r w_r}, \tag{8}$$

where $w_s = 1$ in the reported ranking and w_s is the rescaled DFL weight in the counterfactual ranking. The top X% of a ranking contains the highest-scoring schools up to a total weight of X% of $\sum_r w_r$, so the school at the cutoff counts fractionally.

Figure 3: Non-public Share of the Top of the Ranking, Composite Score



Notes: Share of non-public schools among the top X% of the ranking of internally evaluated schools, by reported score and under the DFL counterfactual ranking described in Section 6.3. The dashed horizontal line is the share of non-public schools among internally evaluated schools (24.7%). Appendix Table B.17 reports the shares in the top 10% and 25% with randomization-inference p-values.

Figure 3 plots the non-public share of the top X% of each ranking, for X between 5 and 50. Under internal evaluation, non-public schools are 25% of schools but 54% of the top decile and 52% of the top quartile. Under external grading, their share falls to 16% of the top decile and 32% of the top quartile. The gap between the two rankings is largest at the very top and

narrows as the top group widens: non-public schools account for about 70% of the reported top 5% but only about 13% of the counterfactual top 5%, whereas the two shares are close by the top half of the ranking. The average percentile rank of non-public schools also falls by 8.6 points under external grading. Randomization inference indicates that these shifts are unlikely to arise from chance variation in which schools were evaluated externally, particularly at the top of the ranking. Appendix Table B.17 reports these statistics and shows that the point estimates are similar when the propensity score also includes school composition, and that the change in the top quartile is similar without reweighting.

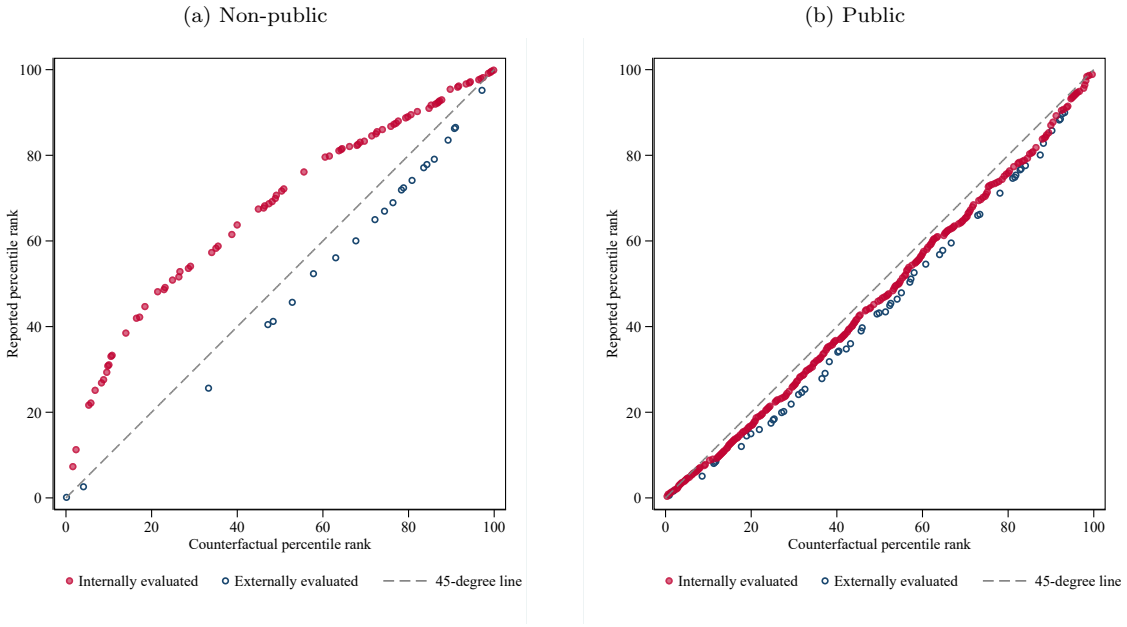
School-level counterfactual ranking. The counterfactual distribution does not tell us which schools a ranking promotes. We therefore construct a counterfactual score for each school. We do this by estimating Equation (6) with the school's four 2016 subject scores, when all schools were externally evaluated, as additional controls, and subtracting the estimated effect of internal evaluation from the score of each internally evaluated school. The counterfactual score of school i is

$$\tilde{Y}_i = Y_i + (1 - \text{External}_i)\,(\hat{\beta}_1 + \hat{\beta}_3\,\text{Non-public}_i), \tag{9}$$

where $\hat{\beta}_1$ and $\hat{\beta}_3$ are the estimates from this regression, so that $-(\hat{\beta}_1 + \hat{\beta}_3\,\text{Non-public}_i)$ is the estimated effect of internal evaluation. Externally evaluated schools keep their reported score, and we re-rank all schools by $\tilde{Y}_i$. Because the same amount is subtracted from all internally evaluated schools of a given ownership type, this exercise does not reorder schools within an ownership type, as in our conceptual framework.

Figure 4 plots each school's reported percentile rank against its counterfactual rank. Internal evaluation raises the composite score of non-public schools by 32 points. In panel (a), we find that every internally evaluated non-public school lies on or above the 45-degree line. On average, these schools gain 14.9 percentile points, close to the estimate in Table 6. As the framework predicts, the gains are largest in the middle of the distribution, where schools are most densely packed: the median internally evaluated non-public school gains 16.3 points, while schools near the top gain little. In panel (b), we find that public schools lie slightly below the 45-degree line and lose 3.6 percentile points on average. At the top, the reported ranking places 9 additional non-public schools in the top decile, and 22.5% of the schools in the reported top decile would not be there under external evaluation. Randomization inference indicates that these changes are unlikely to arise by chance. Appendix Table B.18 reports these statistics and shows that the

Figure 4: Reported and Counterfactual Percentile Ranks, Composite Score



Notes: Each marker is a school. Percentile ranks among the 404 schools of the main sample with valid 2016 scores. Counterfactual ranks use the school-level counterfactual score in Equation (9), as in Appendix Table B.18, column (1). Schools above the 45-degree line rank higher than under the counterfactual.

results are similar when the effect varies with the 2016 score and when we rank all non-Capital schools.

Comparing the two counterfactual rankings. The DFL and school-level counterfactuals give different answers at the very top of the ranking because they impose different assumptions. The school-level counterfactual subtracts the same estimated effect of internal evaluation, 32 points, from every internally evaluated non-public school, so it keeps the order of these schools and shifts all of them by the same amount. Under this counterfactual, the non-public share of the top decile falls from 55% to 38% (Appendix Table B.18). The DFL counterfactual instead uses the whole score distribution of externally evaluated schools, so the effect of internal evaluation can differ along the distribution. Under this counterfactual, the share falls from 54% to 16%. The quantile regressions show that internal evaluation raises non-public scores more in the upper half of the distribution (44 and 46 points at the 75th and 90th percentiles, against 30 at the median), so a constant shift understates how far the top of the non-public distribution moves, and the school-level counterfactual gives the smaller change. In exchange, the DFL counterfactual rests on only 22 externally evaluated non-public schools and describes distributions rather than individual schools. The two exercises also rank slightly different sets of schools: the DFL counterfactual ranks the 328 internally evaluated schools, and the school-level

counterfactual ranks all 404 schools with 2016 scores. We therefore read them as a range: under external grading, the non-public share of the top decile would fall from about 55% to between 16% and 38%.

By subject, the rank effects are largest in Mathematics in all three exercises (Appendix Tables B.19, B.20 and B.21, and Appendix Figures A.6, A.7 and A.8). Taken together, the three exercises show that internal evaluation moves non-public schools up the ranking and over-represents them at the top, where rankings matter most.

7 Conclusions

This paper studies how the mode of evaluation, internal versus external, interacts with incentives to shape measured performance in standardized exams. Using a large-scale randomized controlled trial covering all primary schools in four of Madrid's five school districts, we provide causal evidence that external evaluation lowers test scores, with effects that are both statistically and economically meaningful. Crucially, these declines are not uniform: they are concentrated among non-public schools, which face stronger incentives to appear high-performing.

A simple interpretation of these results emerges from an incentives-based perspective. When schools are evaluated using metrics that matter for reputation, demand, or future outcomes, and when they retain discretion over how those metrics are implemented, they have an incentive to influence measured performance. Our findings show that this incentive translates into systematically higher scores under internal evaluation, particularly among non-public schools which are more exposed to competitive pressures. When monitoring is tightened through external evaluation, the scope for such behavior is reduced, leading to lower measured outcomes. Because these incentives differ across school types, manipulation also distorts rankings, and external evaluation should lower the position of non-public schools and reduce their presence at the top of the ranking.

The rest of our evidence is consistent with this interpretation. External evaluation increases non-participation in non-public schools relative to public schools, so selective participation is one relevant margin, but bounds that accommodate differential selection into test-taking show that it cannot account for the full ownership gradient. Similarly, explanations based on testing conditions or on students' responses to external proctors would need to be systematically stronger in non-public schools. The distributional evidence further shows that internal evalu-

ation moves non-public schools out of the middle of the score distribution and into its upper half. As a result, and as the model predicts, it distorts school rankings: external evaluation lowers the rank of non-public schools by 16 percentile points, and under external grading their share of the top decile would fall from about 55% to between 16% and 38%, depending on the counterfactual.

These findings have important implications for how performance metrics should be interpreted and used. In settings where incentives are heterogeneous, allowing agents discretion over measurement can lead to systematic distortions that vary across institutions. In our context, this implies that internally evaluated standardized exams may not provide a reliable basis for comparing or ranking schools, as differences in scores may reflect not only differences in learning but also differences in incentives to inflate performance. These distortions are largest at the top of the ranking, which is likely the most salient part for families and administrators. External monitoring mitigates these distortions by aligning measured outcomes more closely with underlying achievement.

More broadly, the paper highlights a general lesson: the informational content of performance metrics depends critically on the interaction between incentives and monitoring. When agents have strong incentives to appear high-performing and sufficient discretion over how performance is measured, observed outcomes are likely to reflect a combination of true performance and strategic behavior. Designing measurement systems that account for this interaction—either through external monitoring or alternative institutional arrangements—is therefore essential for ensuring that such metrics serve as credible signals for decision-making.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge (2023) "When Should You Adjust Standard Errors for Clustering?" *The Quarterly Journal of Economics*, 138 (1), 1–35, 10.1093/qje/qjac038. (return to page 18)

Angrist, Joshua D, Erich Battistin, and Daniela Vuri (2017) "In a small moment: Class size and moral hazard in the Italian Mezzogiorno," *American Economic Journal: Applied Economics*, 9 (4), 216–49, 10.1257/app.20160267. (return to page 27)

Bertoni, Marco, Giorgio Brunello, and Lorenzo Rocco (2013) "When the cat is near, the mice won't play: The effect of external examiners in Italian schools," *Journal of Public Economics*, 104, 65–77, 10.1016/j.jpubeco.2013.04.010. (return to page 5, 27)

Borcan, Oana, Mikael Lindahl, and Andreea Mitrut (2017) "Fighting Corruption in Education: What Works and Who Benefits?" *American Economic Journal: Economic Policy*, 9 (1), 180–209, 10.1257/pol.20150074. (return to page 5)

Cai, Xiqian, Yi Lu, Jessica Pan, and Songfa Zhong (2019) "Gender gap under pressure: evidence from China's National College entrance examination," *Review of Economics and Statistics*, 101 (2), 249–263, 10.1162/rest_a_00749. (return to page 27)

Chen, Yijuan and Peter Sivey (2021) "Hospital report cards: Quality competition and patient selection," *Journal of health economics*, 78, 102484, 10.1016/j.jhealeco.2021.102484. (return to page 2)

Chen, Yiqun and Marcus Dillender (2025) "Government Monitoring of Health Care Quality: Evidence from the Nursing Home Sector," Working Paper 34037, National Bureau of Economic Research, https://www.nber.org/papers/w34037. (return to page 2, 5)

Cullen, Julie Berry and Randall Reback (2006) "Tinkering toward accolades: School gaming under a performance accountability system," in Gronberg, Timothy J. and Dennis W. Jansen eds. *Improving School Accountability*, 14 of Advances in Applied Microeconomics, 1–34: Emerald Group Publishing, 10.1016/S0278-0984(06)14001-8. (return to page 28)

Dee, Thomas S., Will Dobbie, Brian A. Jacob, and Jonah Rockoff (2019) "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations," *American Economic Journal: Applied Economics*, 11 (3), 382–423, 10.1257/app.20170520. (return to page 5, 27)

DiNardo, John, Nicole M Fortin, and Thomas Lemieux (1996) "Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach," *Econometrica*, 64 (5), 1001–1044, 10.2307/2171954. (return to page 4, 30, 70, 71)

Figlio, David N. (2006) "Testing, crime and punishment," *Journal of Public Economics*, 90 (4-5), 837–851, 10.1016/j.jpubeco.2005.01.003. (return to page 28)

Figlio, David N and Lawrence S Getzler (2006) "Accountability, ability and disability: Gaming the system?" in Gronberg, Timothy J. and Dennis W. Jansen eds. *Improving School Accountability*, 14 of Advances in Applied Microeconomics, 35–49: Emerald Group Publishing, 10.1016/S0278-0984(06)14002-X. (return to page 27, 28)

Firpo, Sergio, Nicole M Fortin, and Thomas Lemieux (2009) "Unconditional quantile regressions," *Econometrica*, 77 (3), 953–973, 10.3982/ECTA6822. (return to page 30)

Gortázar, Lucas, David Mayor, and José Montalbán (2023) "Residence-based priorities and school choice," *Economics of Education Review*, 95, 102384, 10.1016/j.econedurev.2023.102384. (return to page 7)

Greaves, Ellen and Iftikhar Hussain (2026) "The importance of school quality ratings for school choices: Evidence from a nationwide system," *Journal of Public Economics*, 263, 105787, 10.1016/j.jpubeco.2026.105787. (return to page 2, 5)

Hastings, Justine S and Jeffrey M Weinstein (2008) "Information, school choice, and academic achievement: Evidence from two experiments," *The Quarterly journal of economics*, 123 (4), 1373–1414, 10.1162/qjec.2008.123.4.1373. (return to page 2, 5)

Imbens, Guido W. and Charles F. Manski (2004) "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72 (6), 1845–1857, 10.1111/j.1468-0262.2004.00555.x. (return to page 56)

Imbens, Guido W and Jeffrey M Wooldridge (2009) "Recent developments in the econometrics of program evaluation," *Journal of economic literature*, 47 (1), 5–86, 10.1257/jel.47.1.5. (return to page 24)

Jacob, Brian A. (2005) "Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools," *Journal of Public Economics*, 89 (5-6), 761–796, 10.1016/j.jpubeco.2004.08.004. (return to page 28)

Jacob, Brian A and Steven D Levitt (2003) "Rotten apples: An investigation of the prevalence and predictors of teacher cheating," *The Quarterly Journal of Economics*, 118 (3), 843–877, 10.1162/00335530360698441. (return to page 5, 27)

Kane, Thomas J. and Douglas O. Staiger (2002) "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives*, 16 (4), 91–114, 10.1257/089533002320950993. (return to page 5)

Lavy, Victor (2009) "Performance pay and teachers' effort, productivity, and grading ethics," *American Economic Review*, 99 (5), 1979–2011, 10.1257/aer.99.5.1979. (return to page 2)

Lee, David S. (2009) "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76 (3), 1071–1102, 10.1111/j.1467-937X.2009.00536.x. (return to page 4, 29, 56)

Lemke, Robert J, Claus M Hoerandner, and Robert E McMahon (2006) "Student Assessments, Non-test-takers, and School Accountability," *Education Economics*, 14 (2), 235–250, 10.1080/09645290600622970. (return to page 28)

Lucifora, Claudio and Marco Tonello (2015) "Cheating and social interactions. Evidence from a randomized experiment in a national evaluation program," *Journal of Economic Behavior & Organization*, 115, 45–66, 10.1016/j.jebo.2014.12.006. (return to page 27)

Martinelli, César, Susan W. Parker, Ana Cristina Pérez-Gea, and Rodimiro Rodrigo (2018) "Cheating and Incentives: Learning from a Policy Experiment," *American Economic Journal: Economic Policy*, 10 (1), 298–325, 10.1257/pol.20150066. (return to page 5)

Montalbán-Castilla, José and Almudena Sevilla (2019) "The gender gap in student performance: The role of test-taking environment," *Manuscript*. (return to page 20)

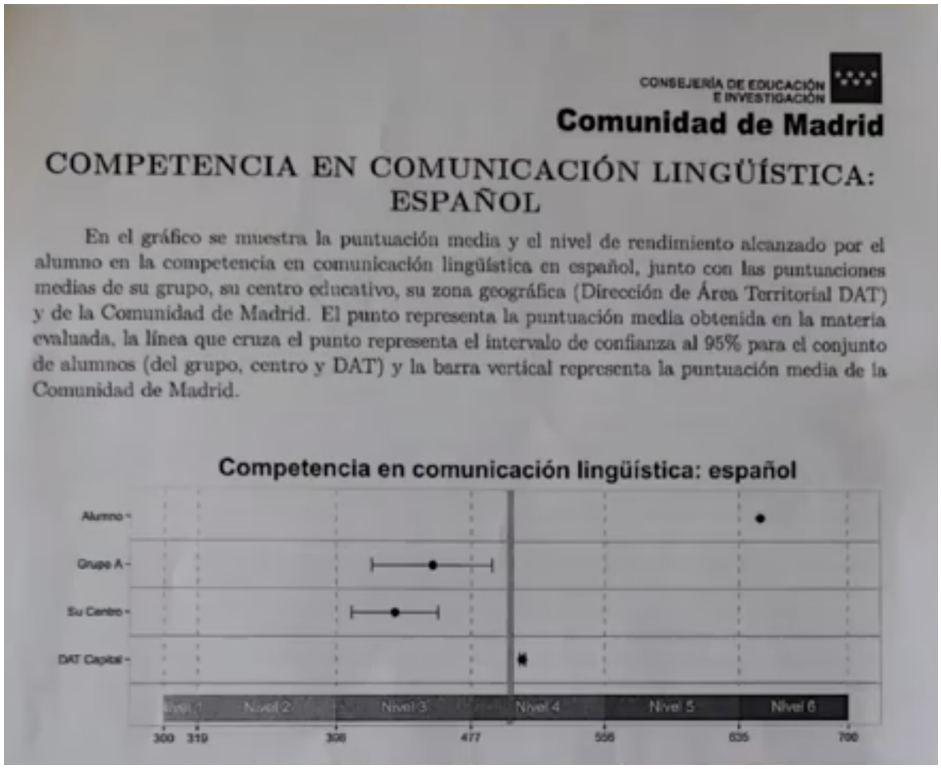
Schlosser, Analia, Zvika Neeman, and Yigal Attali (2019) "Differential Performance in High versus Low Stakes Tests: Evidence from the GRE Test," *The Economic Journal*, 129 (623), 2916–2948, 10.1093/ej/uez015. (return to page 28)

Tyrefors Hinnerich, Björn and Jonas Vlachos (2017) "The Impact of Upper-Secondary Voucher School Attendance on Student Achievement. Swedish Evidence Using External and Internal Evaluations," *Labour Economics*, 47, 1–14, 10.1016/j.labeco.2017.03.009. (return to page 5)

Zou, Eric Yongchen (2021) "Unwatched pollution: The effect of intermittent monitoring on air quality," *American Economic Review*, 111 (7), 2101–2126, 10.1257/aer.20181346. (return to page 2, 5)

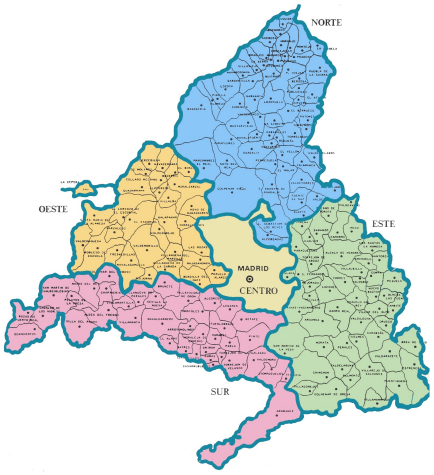
A Additional Figures

Figure A.1: Scorecard to Parents in the LOMCE Exams



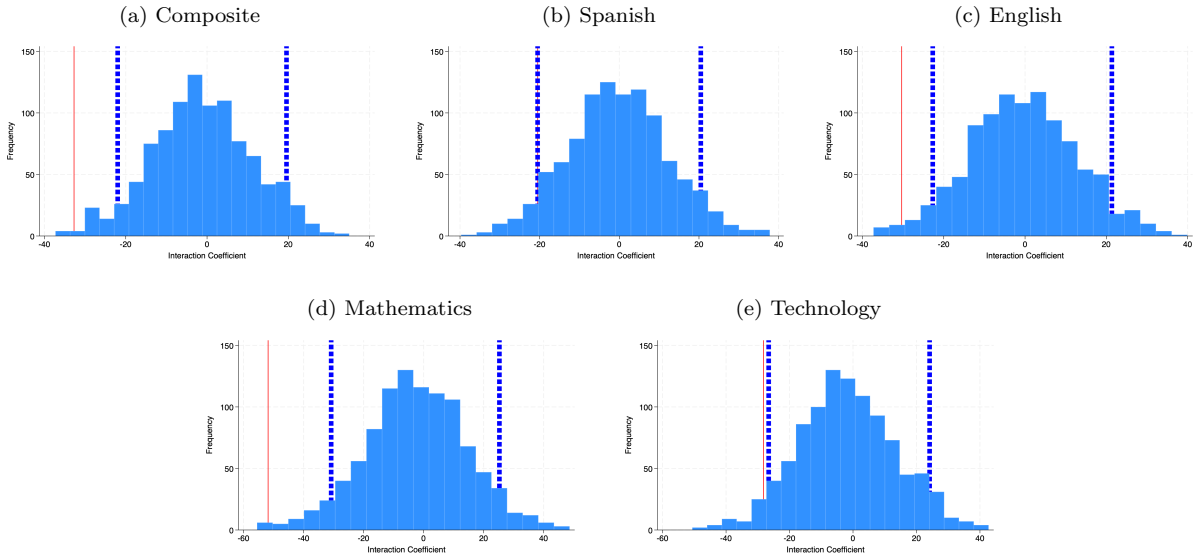
Notes: Scores are divided into six categories, with higher values in the horizontal axis indicating better performance. The first row depicts the performance of the student, the second the average performance of its class, the third the average performance of the school, and the fourth one the average performance in the school's DAT. 95% confidence intervals are also depicted. The vertical line indicates the average across the whole Community of Madrid.

Figure A.2: Map of Madrid's Five Districts (DATs)



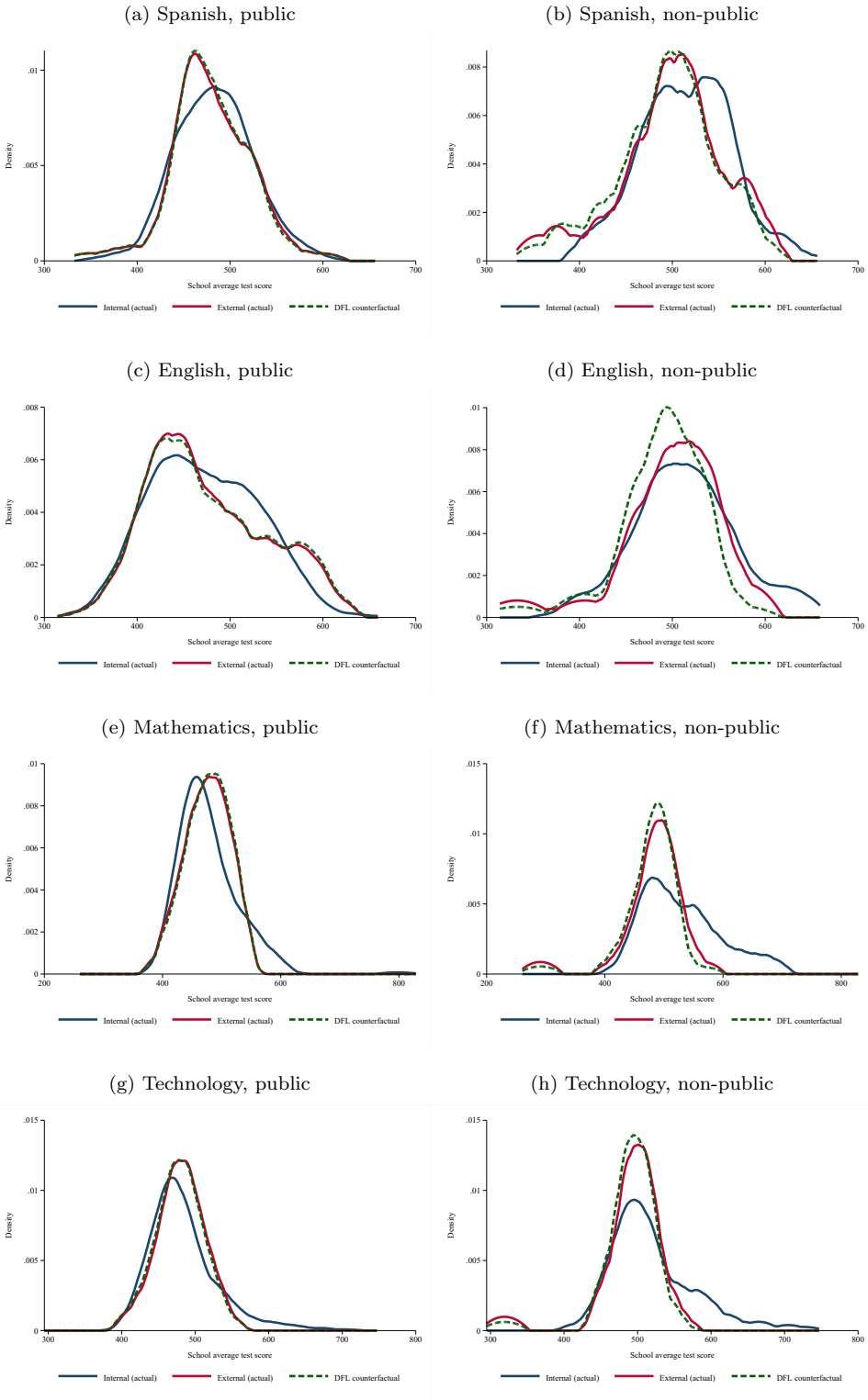
Notes: The five DATs are: North (in blue), East (in green), West (in orange), South (in pink), and Centro in plain yellow.

Figure A.3: Randomization Inference: Interaction Coefficients for Non-public Schools



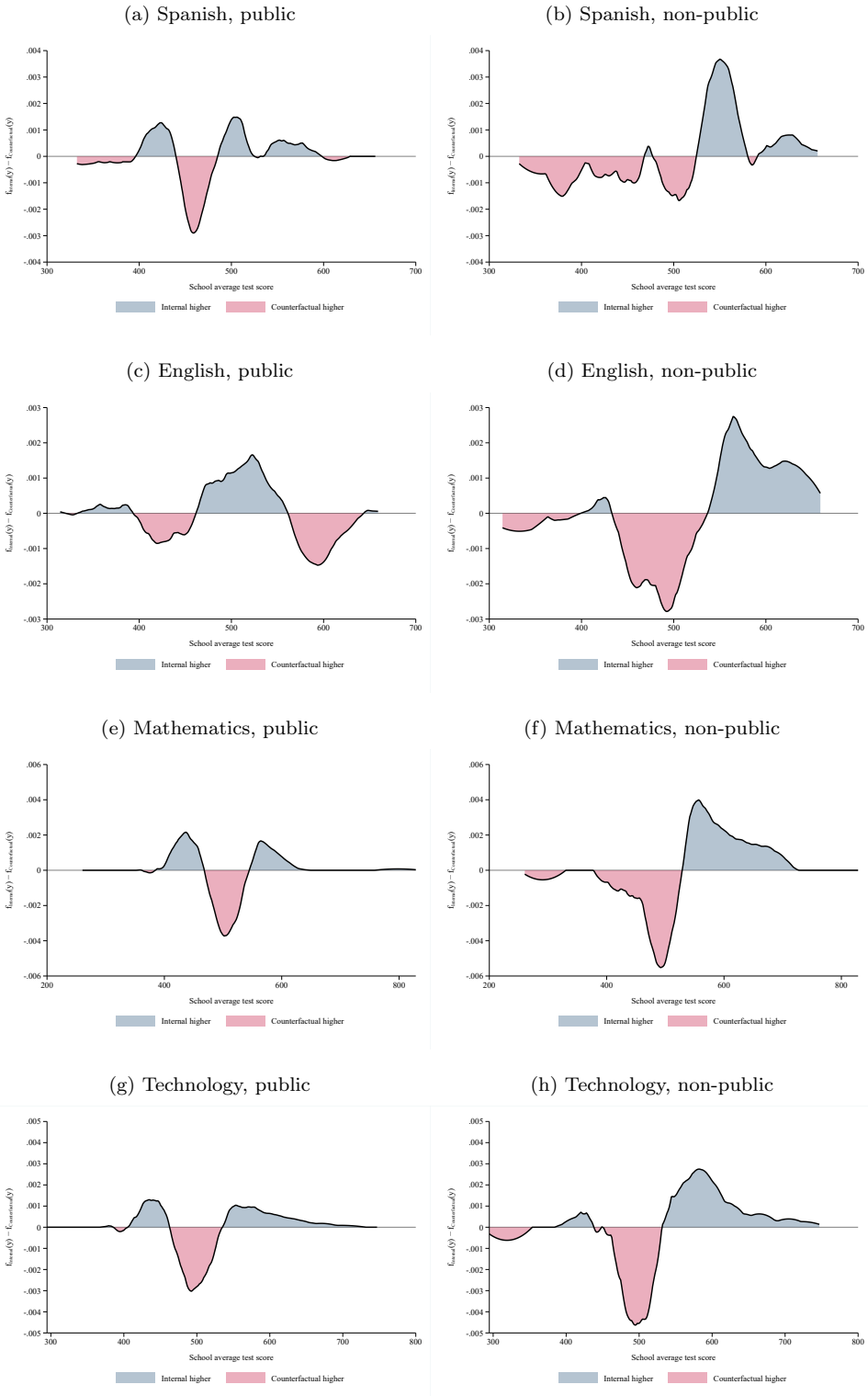
Notes: Histogram of the estimated interaction between external evaluation and the non-public school indicator over 1,000 placebo assignments, in which school-level treatment is randomly reassigned within the randomization strata, preserving the number of treated schools in each stratum. Dashed blue lines indicate the 5th and 95th percentiles of the placebo distribution. The solid red line depicts the estimate under the actual assignment, without additional controls (Tables 3 and 4).

Figure A.4: Actual and Counterfactual Distributions of School Scores, by Subject



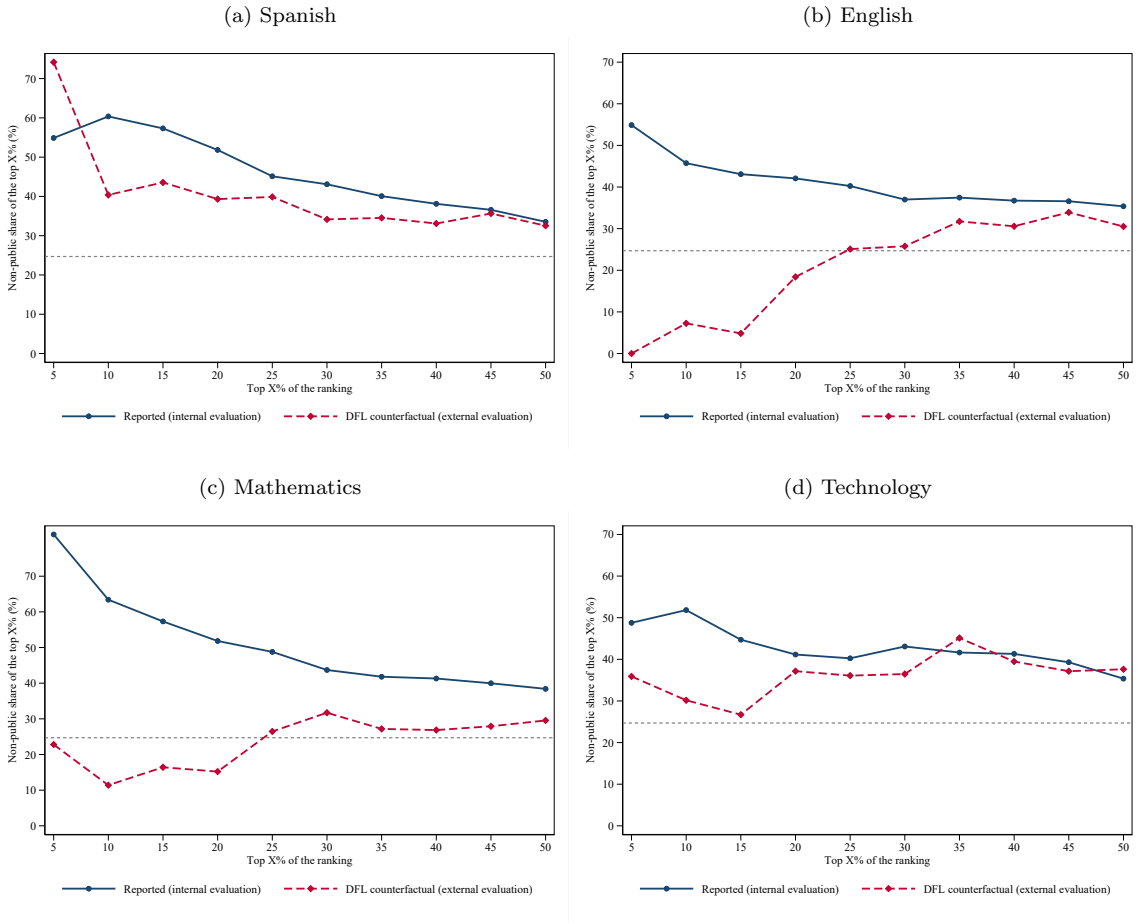
Notes: Kernel densities of the school-level average score in each subject. Internal (actual) and External (actual) are the observed densities under each evaluation regime; the DFL counterfactual is the density of externally evaluated schools reweighted to the observable characteristics of internally evaluated schools of the same ownership type. See the notes to Figure 1 and Appendix E.

Figure A.5: Difference between Actual and Counterfactual Densities, by Subject



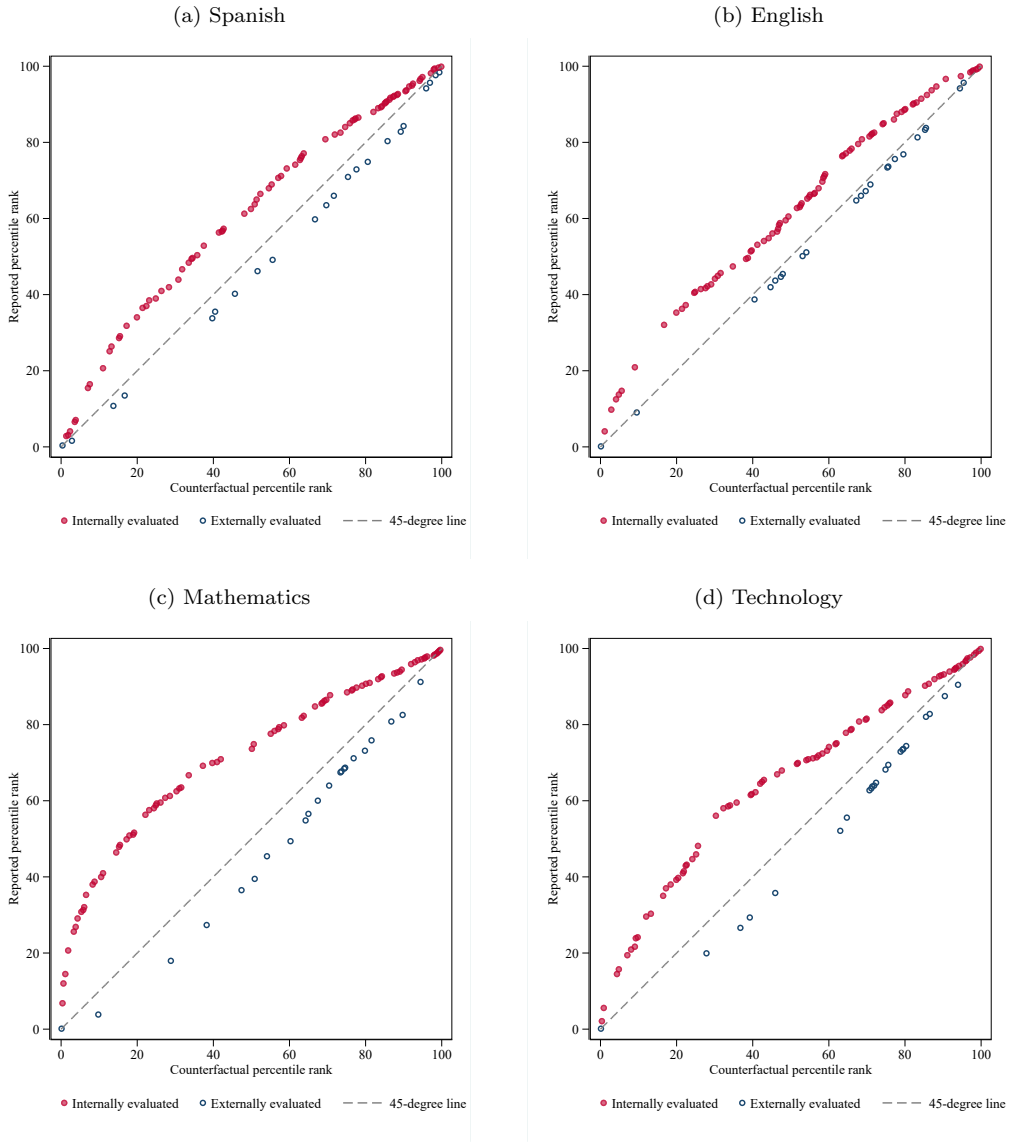
Notes: Each panel plots $\Delta(y) = f_{\text{int}}(y) - f_{\text{cf}}(y)$, the actual density under internal evaluation minus the DFL counterfactual density in Figure A.4. Blue areas mark scores at which internal evaluation places more schools than the counterfactual, and red areas scores at which it places fewer. Within each subject, the public and non-public panels share the same vertical scale.

Figure A.6: Non-public Share of the Top of the Ranking, by Subject



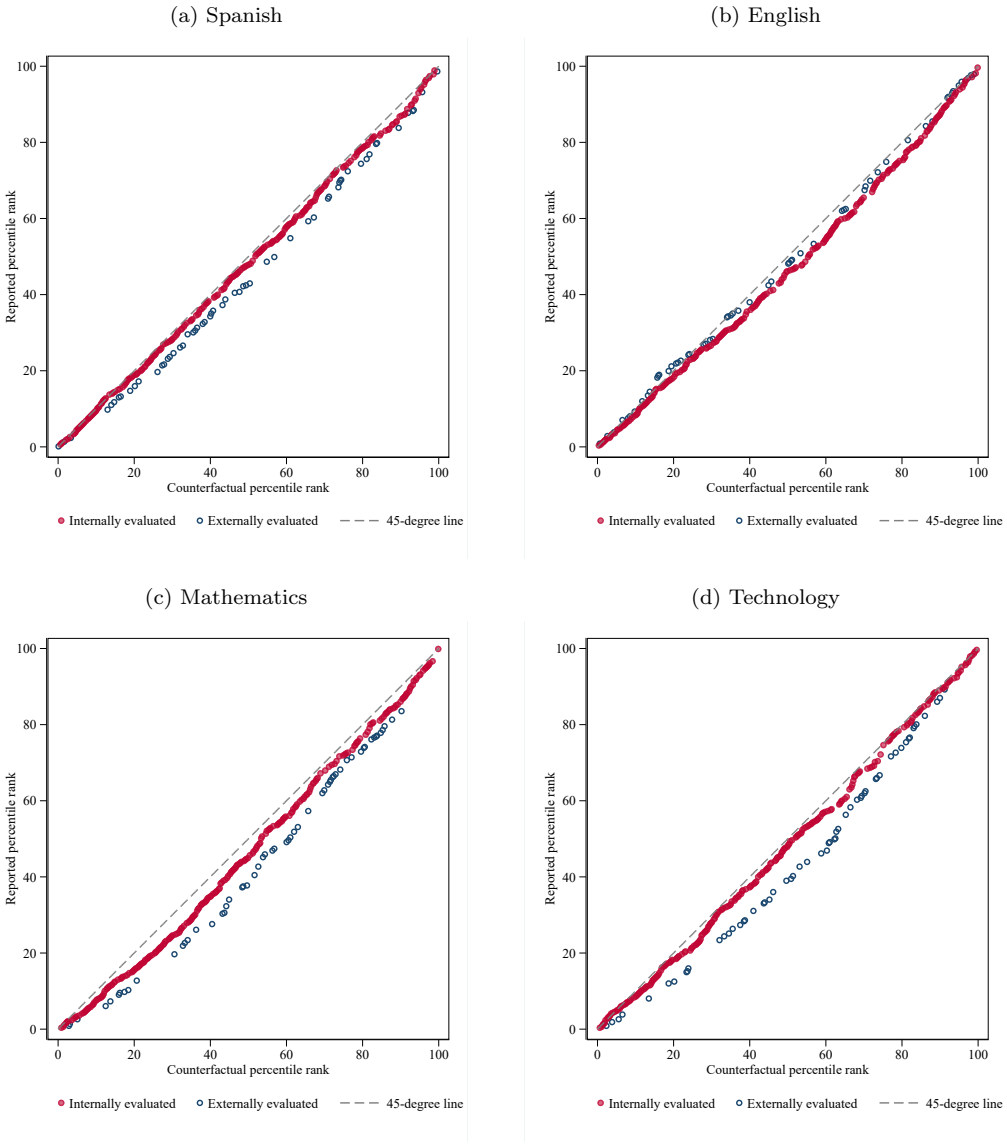
Notes: Share of non-public schools among the top X% of the ranking of internally evaluated schools in each subject, by reported score and under the DFL counterfactual. See the notes to Figure 3.

Figure A.7: Reported and Counterfactual Percentile Ranks of Non-public Schools, by Subject



Notes: See the notes to Figure 4.

Figure A.8: Reported and Counterfactual Percentile Ranks of Public Schools, by Subject



Notes: See the notes to Figure 4.

B Additional Tables

Table B.1: Sample selection, by territorial area

	All	Capital	North	South	East	West
Panel A: Schools						
(1) Valid school identifier	1,295	544	107	322	174	148
(2) Merged with individual questions	1,295	544	107	322	174	148
(3) Non-missing bilingual status	1,274	539	103	317	174	141
(4) Valid scores in all four exams	1,265	537	101	313	173	141
(5) In regression sample	728	0	101	313	173	141
(6) Participation $\geq 30\%$ enrolled	723	0	100	311	172	140
(7) At least 7 students per school	719	0	99	309	171	140
(8) Excluding Capital	719	0	99	309	171	140
(9) Baseline: North and South	408	0	99	309	0	0
Panel B: Students						
(1) Valid school identifier	66,122	27,630	4,792	16,278	9,052	8,370
(2) Merged with individual questions	66,122	27,630	4,792	16,278	9,052	8,370
(3) Non-missing bilingual status	65,065	27,398	4,520	16,096	9,052	7,999
(4) Valid scores in all four exams	59,201	25,063	4,108	14,499	8,117	7,414
(5) In regression sample	34,138	0	4,108	14,499	8,117	7,414
(6) Participation $\geq 30\%$ enrolled	34,118	0	4,105	14,486	8,114	7,413
(7) At least 7 students per school	34,104	0	4,104	14,476	8,111	7,413
(8) Excluding Capital	34,104	0	4,104	14,476	8,111	7,413
(9) Baseline: North and South	18,580	0	4,104	14,476	0	0

Notes: Each row applies the selection criteria cumulatively. Panel A counts schools and Panel B counts students. Columns report the five territorial areas (*Direcciones de Área Territorial*) separately, so that attrition can be read area by area; the territorial restrictions used in the analysis are applied only in rows (8) and (9). Row (5) keeps observations entering the baseline regression, which requires observed treatment assignment and school ownership. Because the Capital district was fully externally evaluated under the rotation scheme, treatment status is undefined there and Capital exits the sample at row (5); row (8) therefore removes no additional observations and is reported for completeness. Row (9) restricts to the North and South districts, where treatment and control groups are balanced on observables, and yields the baseline estimation sample.

Table B.2: Balance Check: All regions

(a) Individual Characteristics

	(1) Female	(2) Non-public	(3) Non-Span.	(4) Parent help	(5) HW freq.	(6) Tablet home	(7) Absences	(8) Repeater
External	-0.012	–	0.005	-0.005	0.004	-0.003	-0.002	-0.022***
	(0.008)	–	(0.010)	(0.014)	(0.004)	(0.009)	(0.003)	(0.007)
Observations	715	719	702	702	702	702	648	702
R-squared	0.013	1.000	0.067	0.014	0.029	0.029	0.053	0.221

(b) Missing Individual Characteristics

	(1) Female miss.	(2) Non-public miss.	(3) Non-Span. miss.	(4) Parent miss.	(5) HW freq. miss.	(6) Tablet miss.	(7) Absences miss.	(8) Repeater miss.
External	0.019	–	0.021	0.020	0.019	0.021	-0.003	0.020
	(0.012)	–	(0.027)	(0.027)	(0.027)	(0.026)	(0.032)	(0.027)
Observations	719		719	719	719	719	719	719
R-squared	0.034		0.028	0.028	0.028	0.028	0.025	0.029

(c) 2016 School Test Scores

	(1) Composite	(2) Spanish	(3) English	(4) Mathematics	(5) Technology
External	6.167**	5.243	5.989**	6.727*	6.709*
	(2.940)	(3.533)	(2.902)	(3.605)	(3.561)
Observations	714	714	714	714	714
R-squared	0.215	0.123	0.519	0.134	0.097

Notes: Replication of the exercise in Table 2 but employing all non-Capital regions as the estimating sample. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.3: Regional Balance of Test Scores

	North					South				
	(1) Comp.	(2) Spanish	(3) English	(4) Math	(5) Tech.	(1) Comp.	(2) Spanish	(3) English	(4) Math	(5) Tech.
External	8.305	3.913	7.340	15.803	6.163	-1.046	3.354	-4.261	-3.456	0.180
	(11.154)	(13.014)	(10.045)	(12.428)	(12.084)	(4.177)	(5.087)	(3.928)	(5.283)	(5.149)
Observations			99					306		
R-squared	0.227	0.138	0.488	0.172	0.124	0.152	0.074	0.478	0.083	0.072
	East					**West**				
	(1) Comp.	(2) Spanish	(3) English	(4) Math	(5) Tech.	(1) Comp.	(2) Spanish	(3) English	(4) Math	(5) Tech.
External	16.115***	9.430	19.425***	20.905***	14.699**	6.960	5.002	9.061	3.504	10.274
	(5.258)	(6.826)	(5.379)	(7.447)	(6.946)	(5.626)	(6.780)	(6.239)	(5.943)	(7.367)
Observations			169					140		
R-squared	0.261	0.118	0.602	0.169	0.077	0.070	0.017	0.279	0.078	0.086

Notes: Replication of the specifications in Table 2's Panel C, estimated separately for the four experimental districts. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.4: Effects of External Grading on Survey Non-Response Rates

	Non-Response Rate for Socio-demographic Survey					
	Without Interactions			With School Type Interactions		
	≥ 2 missing (1)	≥ 3 missing (2)	≥ 4 missing (3)	≥ 2 missing (4)	≥ 3 missing (5)	≥ 4 missing (6)
Main effects						
External	0.001	-0.001	-0.004	-0.011	-0.013	-0.017
	(0.024)	(0.024)	(0.024)	(0.027)	(0.027)	(0.027)
Interaction effects						
Ext. × Charter/Private				0.045	0.046	0.050
				(0.058)	(0.059)	(0.060)
Observations		408			408	
R-squared	0.065	0.065	0.070	0.066	0.067	0.072

Notes: Replication of Table 3 with no controls where the outcomes are the school-level fraction of students in grade 6 that have missing information on at least 2, at least 3, or at least 4 of the following: age enrolled, number of books at home, parental education, parental occupation, and absentism. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.5: Effects of External Grading by Bilingual Status

	(1)	(2)	(3)	(4)
Panel A: Overall Score English				
External	18.887*	-2.511	18.416*	-3.024
	(10.166)	(7.262)	(10.198)	(7.243)
External × Charter/Private	-45.614**	-29.920**	-35.602**	-22.683*
	(17.472)	(13.975)	(16.812)	(13.672)
Sample	Bilingual	Non-bilingual	Bilingual+Non-private	Non-bilingual+Non-private
Observations	126	282	124	267
R-squared	0.230	0.444	0.229	0.389
Panel B: Written English				
External	0.019	-0.077***	0.017	-0.080***
	(0.036)	(0.028)	(0.035)	(0.028)
External × Charter/Private	-0.215***	-0.059	-0.186***	-0.050
	(0.062)	(0.048)	(0.064)	(0.048)
Sample	Bilingual	Non-bilingual	Bilingual+Non-private	Non-bilingual+Non-private
Observations	126	282	124	267
R-squared	0.217	0.364	0.223	0.314

Notes: Replication of the school-level with-controls specification in Panel B of Table 3 for English performance, run separately for schools classified as bilingual and non-bilingual in columns (1) and (2), respectively. Columns (3) and (4) further remove private schools, which are not officially part of the bilingual program of the Community of Madrid. Panel A's outcome is the overall score in English while Panel B's uses the score in the written portion of the English exam as the dependent variable. Robust standard errors clustered at school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.6: Effects of External Grading on Academic Scores, Grade 6 (2017) – Separating Charter and Private Schools

	No Controls					**With Controls**				
	Composite	Spanish	English	Math	Tech.	Composite	Spanish	English	Math	Tech.
Main treatment effect										
External	-0.985	-3.273	4.622	-1.304	-3.987	-1.984	-4.455	5.209	-3.943	-4.747
	(4.441)	(5.850)	(6.025)	(5.510)	(5.021)	(4.018)	(5.328)	(5.763)	(5.514)	(4.941)
Treatment × School type										
External × Charter	-22.860**	-14.006	-20.076*	-42.884***	-14.474	-29.113***	-20.064	-28.432**	-47.777***	-20.179*
	(9.325)	(13.529)	(11.200)	(11.845)	(10.313)	(8.839)	(12.424)	(11.110)	(11.845)	(10.444)
External × Private	-74.343*	-45.214	-72.553	-89.266*	-90.341**	-66.209**	-32.099	-66.787*	-83.323*	-82.626**
	(41.649)	(40.840)	(44.008)	(50.797)	(38.409)	(33.029)	(29.795)	(35.517)	(43.929)	(32.425)
Observations			408					408		
R-squared	0.305	0.215	0.496	0.213	0.140	0.431	0.372	0.579	0.302	0.214

Notes: Regressions follow Equation (6), with separate interactions of external evaluation with indicators for charter and private schools; the omitted category is public schools. The left part does not include additional controls beyond the stratification variables. The right part includes as controls the school-level averages of the variables in Table 2's Panel A. Composite score is the average of the four subject scores. Robust standard errors clustered at school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.7: Heterogeneous Effects of External Grading by School Ownership (Controlling for School's Performance in 2016)

	Without Student Controls					With Student Controls				
	(1) Composite	(2) Spanish	(3) English	(4) Math	(5) Tech.	(1) Composite	(2) Spanish	(3) English	(4) Math	(5) Tech.
Panel A: Private/Concertado vs. Public										
External	-2.826	-4.052	3.806	-4.500	-6.558	-3.592	-5.066	4.270	-6.699	-6.874
	(4.275)	(5.683)	(5.778)	(5.549)	(5.010)	(3.936)	(5.241)	(5.712)	(5.653)	(4.905)
External × Charter/Private	-24.750**	-16.282	-23.950**	-39.659***	-19.109*	-28.663***	-19.259	-29.708***	-43.061***	-22.626**
	(9.969)	(13.243)	(11.461)	(12.656)	(10.739)	(8.973)	(11.703)	(10.916)	(12.134)	(10.548)
Observations			405					405		
R-squared	0.391	0.249	0.539	0.315	0.242	0.496	0.392	0.608	0.383	0.304

Notes: Replication of panel B in Table 3, where we additionally control for the school's average LOMCE score in the subject in the previous academic year. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.8: Heterogeneous Effects of External Grading by School Ownership: Controlling for ISEC

	Without Controls					With Controls				
	(1) Composite	(2) Spanish	(3) English	(4) Math	(5) Tech.	(1) Composite	(2) Spanish	(3) English	(4) Math	(5) Tech.
External	-2.616	-5.179	2.679	-2.790	-5.175	-3.417	-6.235	3.517	-5.284	-5.667
	(4.083)	(5.323)	(5.457)	(5.325)	(5.054)	(3.857)	(5.185)	(5.482)	(5.369)	(4.997)
External × Charter/Private	-29.774***	-16.117	-26.819**	-48.818***	-27.342**	-33.280***	-18.066	-32.408***	-51.974***	-30.673***
	(10.866)	(13.130)	(11.930)	(13.587)	(12.576)	(9.902)	(11.793)	(11.687)	(13.084)	(11.838)
Observations			407					407		
R-squared	0.383	0.318	0.556	0.251	0.165	0.453	0.406	0.597	0.310	0.217

Notes: Replication of Table 3 additionally controlling for the Socioeconomic and Cultural Index (ISEC); This index, defined in PISA tests as ESCS, is computed from information on: i) the highest level of parental education; ii) the highest level of parental occupation; iii) the level of resources in the household (including technological devices, internet connection, etcetera); iv) the number of books in the household. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.9: Heterogeneous Effects of External Grading by School Ownership: Robustness to using average total correct responses

	Without Controls					With Controls				
	(1) Composite	(2) Spanish	(3) English	(4) Math	(5) Tech.	(1) Composite	(2) Spanish	(3) English	(4) Math	(5) Tech.
Panel A: Charter/Private vs. Public										
External	-0.284***	-0.170	-0.508**	-1.040***	-0.631***	-0.309***	-0.227	-0.518**	-1.162***	-0.656***
	(0.090)	(0.211)	(0.215)	(0.281)	(0.143)	(0.082)	(0.180)	(0.213)	(0.276)	(0.137)
External × Charter/Private	-0.592***	-0.903*	-0.871	-2.595***	-0.763**	-0.660***	-1.004**	-1.092**	-2.760***	-0.848***
	(0.221)	(0.460)	(0.545)	(0.657)	(0.340)	(0.193)	(0.405)	(0.484)	(0.605)	(0.307)
Observations			408					408		
R-squared	0.291	0.212	0.522	0.233	0.171	0.411	0.357	0.600	0.318	0.261

Notes: Replication of Panel B of Table 3, where the outcome is the school-average number of correct answers in each subject instead of the official standardized score. The composite averages the four subject counts after standardizing each across schools, so it is measured in school-level standard deviations. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.10: Heterogeneous Effects of External Grading by School Ownership: Individual-level Analysis

	Without Controls					With Controls				
	Comp.	Spanish	English	Math	Tech.	Comp.	Spanish	English	Math	Tech.
Panel A: Charter/Private vs. Public										
External	0.567	-2.050	6.314	2.100	-4.094	-0.359	-2.676	5.587	0.729	-5.078
	(4.052)	(5.634)	(5.597)	(4.964)	(4.700)	(3.648)	(5.405)	(5.271)	(4.650)	(4.351)
External × Charter/Private	-30.389***	-15.805	-33.604***	-50.678***	-21.468**	-31.334***	-17.093	-34.717***	-51.225***	-22.300**
	(8.592)	(11.933)	(9.589)	(11.691)	(10.707)	(8.285)	(11.352)	(9.534)	(11.589)	(10.365)
Observations			18,580					18,580		
R-squared	0.105	0.053	0.184	0.085	0.038	0.209	0.173	0.271	0.147	0.088

Notes: Replication of panels B in Table 3, where the unit of observation is the student instead of the school. Robust standard errors clustered at school level in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.11: Heterogeneous Effects of External Grading by School Ownership: North and South Districts

	Without Controls					With Controls				
	Composite	Spanish	English	Math	Tech.	Composite	Spanish	English	Math	Tech.
North Region										
External	9.235	3.547	18.665*	6.561	8.168	5.702	5.166	13.955	0.228	3.459
	(7.596)	(11.519)	(9.822)	(10.795)	(8.943)	(10.758)	(12.956)	(12.883)	(14.886)	(12.728)
External × Private/Concertado	-45.999	-20.420	-50.824	-64.375*	-48.375	-47.333**	-28.429	-48.480*	-64.366**	-48.058**
	(28.887)	(32.412)	(31.715)	(35.240)	(29.269)	(21.518)	(24.670)	(27.863)	(29.751)	(22.905)
Observations			99					99		
R-squared	0.350	0.218	0.556	0.259	0.185	0.579	0.495	0.730	0.429	0.385
South Region										
External	-4.656	-5.722	-0.422	-4.128	-8.353	-6.309	-7.110	-0.126	-8.089	-9.910*
	(5.316)	(6.775)	(7.303)	(6.377)	(5.947)	(4.607)	(5.988)	(6.951)	(6.267)	(5.819)
External × Private/Concertado	-27.741***	-20.994	-22.553*	-46.996***	-20.419*	-30.636***	-20.864	-28.466**	-50.134***	-23.082*
	(10.157)	(14.266)	(11.949)	(12.672)	(11.727)	(9.625)	(13.478)	(11.600)	(12.661)	(11.808)
Observations			309					309		
R-squared	0.249	0.152	0.431	0.182	0.107	0.374	0.309	0.519	0.274	0.180

Notes: Replication of the specifications in Table 3, estimated separately for the North and South districts. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.12: Heterogeneous Effects of External Grading by School Ownership: All Regions

	Without Controls					With Controls				
	(1) Composite	(2) Spanish	(3) English	(4) Math	(5) Tech.	(1) Composite	(2) Spanish	(3) English	(4) Math	(5) Tech.
Panel A: Charter/Private vs. Public										
External	2.461	3.320	5.132	1.701	-0.311	-0.816	0.750	1.623	-2.510	-3.126
	(3.472)	(4.289)	(4.464)	(4.360)	(4.018)	(3.258)	(4.013)	(4.324)	(4.354)	(4.007)
External × Charter/Private	-23.851***	-18.679**	-17.496*	-34.828***	-24.401***	-23.037***	-17.899**	-16.623*	-33.912***	-23.714***
	(7.443)	(8.965)	(9.229)	(9.276)	(8.413)	(6.873)	(8.508)	(8.576)	(8.995)	(8.099)
Observations			719					719		
R-squared	0.280	0.177	0.479	0.204	0.151	0.412	0.313	0.576	0.294	0.229

Notes: Replication of the specifications in Table 3 but employing all non-Capital regions as the estimating sample. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.13: Effects on Writing Scores in Spanish and English: Disaggregated by School Ownership

	(1) Composite Writing Score
Treatment effect	
External	-0.066***
	(0.015)
Treatment × School type interactions	
External × Charter/Private	-0.070**
	(0.028)
Observations	408
R-squared	0.485

Notes: School-level replication of Panel B of Table 3, with controls, where the outcome is the school average of each student's mean share of the maximum score in the writing sections of the Spanish and English exams. Only students with both writing scores are included. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.14: Effects of External Grading on Student Absences in Grade 6

	Absence Rate (Grade 6)	
	(1)	(2)
Treatment effect		
External grading	-0.009	-0.024**
	(0.009)	(0.010)
Treatment × School type interactions		
External × Charter/Private		0.058***
		(0.020)
Observations	408	408
R-squared	0.185	0.198
Mean Dep. Var.	0.063	0.063

Notes: Column (1) replicates the specification in Table 3's Panel A while column (2) replicates the specification in Panel B. The outcome is the school-level share of students who do not have scores for all subjects out of the total number of enrolled students. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.15: Lee (2009) Bounds for Treatment Effects by Subject and School Type

	Subject				
	Composite	Spanish	English	Math	Tech.
Public schools	[-1.84, 4.81]	[-6.91, 3.30]	[0.07, 11.35]	[0.47, 6.89]	[-9.45, 0.33]
95% CI	[-5.64, 8.89]	[-11.33, 7.50]	[-5.89, 17.19]	[-4.03, 11.07]	[-14.06, 5.31]
Private/Concertado	[-28.78, -20.08]	[-18.81, -7.30]	[-25.90, -16.17]	[-47.68, -37.61]	[-24.36, -14.73]
95% CI	[-33.39, -15.53]	[-24.46, -1.80]	[-31.50, -10.61]	[-53.52, -31.66]	[-30.28, -8.99]
Difference (Pvt-Pub)	[-33.60, -18.24]	[-22.11, -0.39]	[-37.26, -16.24]	[-54.58, -38.07]	[-24.69, -5.28]
Attrition diff (T-C), pooled					
Public	0.015	0.023	0.021	0.012	0.020
Private	-0.018	-0.019	-0.019	-0.019	-0.018
Attrition diff (T-C), range across strata					
Public	[-0.011, 0.033]	[-0.010, 0.043]	[-0.002, 0.035]	[-0.011, 0.024]	[-0.009, 0.040]
Private	[-0.045, 0.015]	[-0.047, 0.004]	[-0.044, 0.009]	[-0.047, 0.022]	[-0.037, 0.002]

Notes: Lee (2009) bounds for treatment effects on exam scores, accounting for differential attrition. Because treatment was randomized within strata defined by district, bilingual status, and ownership, bounds are computed *within* each stratum and aggregated across strata; trimming proportions and the direction of trimming are therefore stratum-specific. This matters here: as the last two rows show, the sign of differential attrition flips across strata in every subject and ownership group, so pooling would trim the wrong arm in some cells. "Attrition diff" is the difference in response rates between treated and control groups (positive = treatment increases response). 95% confidence intervals for the treatment effect follow Imbens and Manski (2004). Bounds for the difference between school types are obtained as $[\underline{\beta}_{Pvt} - \overline{\beta}_{Pub},\ \overline{\beta}_{Pvt} - \underline{\beta}_{Pub}]$ and are therefore conservative. Confidence intervals are analytic; clustering the bootstrap at the school level (the unit of randomization) widens them substantially.

Table B.16: Quantile Treatment Effects by Subject

(a) Spanish

	(1) q10	(2) q25	(3) q50	(4) q75	(5) q90
External	3.321	-2.916	-3.995	-0.384	-7.646
	(7.443)	(6.400)	(7.254)	(6.239)	(6.497)
Ext. × Priv./Conc.	-35.508*	-28.133	-27.807	-23.750	-8.390
	(18.820)	(19.166)	(20.987)	(18.493)	(14.799)
Pseudo R-squared	0.252	0.204	0.198	0.237	0.300

(b) English

	(1) q10	(2) q25	(3) q50	(4) q75	(5) q90
External	5.526	1.030	4.570	6.541	6.475
	(7.803)	(8.677)	(9.799)	(8.349)	(8.510)
Ext. × Priv./Conc.	-16.403	-32.477*	-36.940**	-43.973**	-58.410***
	(21.513)	(17.468)	(16.892)	(21.777)	(18.003)
Pseudo R-squared	0.403	0.400	0.397	0.370	0.352

(c) Mathematics

	(1) q10	(2) q25	(3) q50	(4) q75	(5) q90
External	6.710	13.122**	4.560	-8.798	-27.789***
	(6.596)	(5.681)	(5.179)	(10.773)	(10.594)
Ext. × Priv./Conc.	-37.259**	-38.355***	-45.668***	-59.413***	-83.202***
	(17.550)	(11.555)	(12.290)	(22.859)	(31.824)
Pseudo R-squared	0.211	0.189	0.163	0.199	0.262

(d) Technology

	(1) q10	(2) q25	(3) q50	(4) q75	(5) q90
External	11.040*	5.636	6.725	-7.262	-30.774***
	(5.758)	(4.970)	(6.850)	(7.782)	(9.252)
Ext. × Priv./Conc.	-14.324	-20.975	-18.949	-23.597*	-34.178
	(14.770)	(18.060)	(12.203)	(13.407)	(21.923)
Pseudo R-squared	0.191	0.150	0.135	0.138	0.197

Notes: Quantile regression estimates at indicated percentiles. All regressions include the same controls as in Table 3. All panels are estimated on 408 observations. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.17: Reported and Counterfactual Rankings of Internally Evaluated Schools, Composite Score

		DFL counterfactual		Difference, alternative counterfactuals	
	Reported	Counterfactual	Difference	Propensity with school composition	Unweighted
	(1)	(2)	(3)	(4)	(5)
Non-public share of the top 10%	54.3	15.8	−38.4	−39.7	−19.2
			[0.035]	[0.053]	[0.271]
Non-public share of the top 25%	52.4	32.3	−20.1	−21.4	−20.0
			[0.045]	[0.065]	[0.030]
Mean percentile rank of non-public schools	69.1	60.5	−8.6	−8.1	−5.7
			[0.096]	[0.168]	[0.220]

Notes: See Section 6.3 and Figure 3. Column (1): ranking of the 328 internally evaluated schools by composite score. Column (2): ranking under the DFL counterfactual, which pools externally evaluated schools reweighted to the design variables of internally evaluated schools of the same ownership type (Appendix E), with each ownership type carrying its share among internally evaluated schools. Column (4) adds the school-composition controls of Table E.1 to the propensity score; column (5) does not reweight. Non-public schools are 24.7% of internally evaluated schools. Because this ranking includes only internally evaluated schools, the mean percentile rank of non-public schools in column (1) differs from the mean in Table 6, which ranks all 408 schools. Shares of the top X% treat the school at the cutoff fractionally. Randomization-inference p-values in brackets.

Table B.18: School-level Counterfactual Ranking, Composite Score

	Constant effect	Effect varies with 2016 score	Constant, non-Capital schools
	(1)	(2)	(3)
Effect of internal evaluation, non-public (points)	32.0	26.7	32.0
	[<0.001]	[0.001]	
Effect of internal evaluation, public (points)	3.6	3.4	3.6
	[0.443]	[0.482]	
Mean rank gain, internally evaluated non-public (pp)	14.9	12.3	14.7
	[0.001]	[0.005]	
Mean rank change, public (pp)	−3.6	−3.0	−3.7
	[0.005]	[0.014]	
Additional non-public schools in the top quartile	15	12	33
	[0.001]	[0.008]	
Additional non-public schools in the top decile	9	9	18
	[0.004]	[0.003]	
Reported top-decile schools not in the counterfactual top decile (%)	22.5	22.5	25.4
	[0.007]	[0.010]	
Non-public share of the top decile, reported / counterfactual (%)	55.0 / 37.5	55.0 / 37.5	60.6 / 40.8
	[0.039]	[0.039]	
Rank gain across internally evaluated non-public schools (pp)			
10th percentile	2.7	1.2	2.2
Median	16.3	12.4	15.3
90th percentile	25.0	21.8	24.0
Schools	404	404	713

Notes: See Section 6.3 and Equation (9). The counterfactual score of an internally evaluated school subtracts the effect of internal evaluation estimated from Equation (6) with the controls of Table 3 and the school's four 2016 subject scores; externally evaluated schools keep their score. In column (2), the effect varies linearly with the school's 2016 score. Column (3) applies the effect estimated in column (1) to all non-Capital schools. Columns (1) and (2) include the schools of the main sample with valid 2016 scores. Rank changes are the reported minus the counterfactual percentile rank. Additional schools: internally evaluated non-public schools in the reported top group minus those in the counterfactual top group. Randomization-inference p-values in brackets (one-sided for the number of additional schools and the share of top-decile schools); not computed in column (3), where only main-sample schools are re-assigned.

Table B.19: Effect of External Evaluation on School Rank, by Subject

	Composite	Spanish	English	Mathematics	Technology
Panel A: Percentile rank					
External × Priv./Conc.	−15.28	−11.06	−12.71	−21.61	−11.34
	(5.82)	(7.22)	(5.31)	(6.26)	(6.46)
	[0.021]	[0.118]	[0.028]	[0.004]	[0.121]
External + External × Priv./Conc.	−15.57	−13.73	−10.90	−19.88	−10.39
	(4.90)	(6.26)	(4.47)	(5.23)	(5.38)
	[0.008]	[0.033]	[0.039]	[0.002]	[0.107]
Panel B: Top quartile					
External × Priv./Conc.	−35.85	−22.78	−25.97	−36.60	−25.66
	(10.82)	(11.36)	(9.83)	(10.58)	(10.31)
	[0.001]	[0.058]	[0.022]	[0.009]	[0.041]
External + External × Priv./Conc.	−34.00	−26.20	−22.32	−42.38	−32.00
	(9.54)	(10.06)	(8.62)	(9.08)	(8.95)
	[<0.001]	[0.018]	[0.024]	[<0.001]	[0.004]
Panel C: Top decile					
External × Priv./Conc.	−20.31	−14.65	−27.74	−22.17	−17.25
	(6.97)	(9.66)	(7.87)	(7.04)	(6.31)
	[0.045]	[0.188]	[<0.001]	[0.032]	[0.110]
External + External × Priv./Conc.	−28.35	−17.63	−17.24	−29.29	−25.85
	(6.77)	(9.24)	(6.59)	(6.95)	(6.09)
	[0.003]	[0.098]	[0.030]	[0.003]	[0.009]

Notes: See the notes to Table 6. Each column ranks the 408 schools by their average score in the indicated subject. Standard errors clustered at the school level in parentheses. Randomization-inference p-values in brackets.

Table B.20: Reported and Counterfactual Rankings of Internally Evaluated Schools, by Subject

	Composite	Spanish	English	Mathematics	Technology
Panel A: Non-public share of the top 10%					
Reported	54.3	60.4	45.7	63.4	51.8
Counterfactual	15.8	40.4	7.3	11.4	30.2
Difference	−38.4	−20.0	−38.5	−52.0	−21.7
	[0.035]	[0.243]	[0.013]	[0.007]	[0.244]
Panel B: Non-public share of the top 25%					
Reported	52.4	45.1	40.2	48.8	40.2
Counterfactual	32.3	39.9	25.1	26.5	36.1
Difference	−20.1	−5.3	−15.1	−22.3	−4.2
	[0.045]	[0.623]	[0.081]	[0.026]	[0.692]
Panel C: Mean percentile rank of non-public schools					
Reported	69.1	64.6	65.1	69.0	65.7
Counterfactual	60.5	58.6	58.1	52.9	62.3
Difference	−8.6	−5.9	−7.0	−16.1	−3.5
	[0.096]	[0.332]	[0.112]	[0.009]	[0.526]

Notes: See the notes to Table B.17, columns (1)–(3). Randomization-inference p-values for the difference in brackets.

Table B.21: School-level Counterfactual Ranking, by Subject

	Composite	Spanish	English	Mathematics	Technology
Effect of internal evaluation, non-public (points)	32.0	23.9	25.3	49.6	29.1
	[<0.001]	[0.020]	[0.010]	[<0.001]	[0.020]
Effect of internal evaluation, public (points)	3.6	5.1	−4.3	6.7	6.9
	[0.443]	[0.322]	[0.467]	[0.306]	[0.257]
Mean rank gain, internally evaluated non-public (pp)	14.9	9.2	10.2	19.0	12.7
	[0.001]	[0.046]	[0.004]	[0.002]	[0.057]
Mean rank change, public (pp)	−3.6	−2.2	−2.6	−4.6	−2.9
	[0.005]	[0.075]	[0.007]	[0.012]	[0.093]
Additional non-public schools in the top quartile	15	8	13	13	9
	[0.001]	[0.030]	[0.006]	[0.012]	[0.037]
Additional non-public schools in the top decile	9	9	6	10	5
	[0.004]	[0.071]	[0.019]	[0.004]	[0.063]
Reported top-decile schools not in the counterfactual top decile (%)	22.5	22.5	15.0	25.0	12.5
	[0.007]	[0.072]	[0.048]	[0.004]	[0.085]
Non-public share of the top decile, reported / counterfactual (%)	55.0 / 37.5	65.0 / 42.5	40.0 / 25.0	62.5 / 37.5	52.5 / 42.5
	[0.039]	[0.044]	[0.026]	[0.004]	[0.151]

Notes: See the notes to Table B.18, column (1). Randomization-inference p-values in brackets.

C Analysis for Grade 10

The 2017 evaluation wave also covered Grade 10 (4th of ESO), with the same experimental design as the Grade 6 exams: the Capital district was fully externally evaluated under the rotation scheme, and schools in the remaining four districts were assigned to external or internal evaluation. The Grade 10 data contain four subjects—Spanish, English, mathematics, and social and civic competences—standardized to mean 500 and standard deviation 100, and record school ownership only as a binary indicator, so non-public pools charter and private schools as in our main analysis. We report these results in the Appendix rather than in the main text because the Grade 10 score records are markedly less complete than the Grade 6 ones, in ways that are correlated with both ownership and the evaluation regime; we first document this and then present the results.

Incomplete score records. The quality of the score data is markedly worse in Grade 10 than in Grade 6: many students in the administrative file have scores in only some of the subjects, or in none. Table C.1 shows the completeness of score records among the students appearing in the file. Two features stand out. *First*, incompleteness is severe and strongly related to ownership: only 70–72 percent of public-school students have valid scores in all four subjects, against 92–94 percent in non-public schools; roughly one in five public-school students has no score at all. *Second*, and more importantly for our design, *partial* records—students with scores in some but not all subjects—are not balanced across evaluation regimes within public schools (5.9 percent under internal versus 9.8 percent under external evaluation), and these partial records exhibit extreme scores: among public-school students evaluated externally, the mean Spanish score of partial records is 201, three student-level standard deviations below the mean, against 500 for complete records. Moreover, these students were mostly not flagged as being in a special situation, which leaves their missing scores unexplained: fewer than 5 percent of partial records carry a special-situation code. Because these low-scoring partial records are concentrated in externally evaluated public schools, and because they plausibly reflect behavioral test-completion responses that we cannot separate from measurement problems with these data, we treat the Grade 10 estimates as complementary evidence rather than as part of the main analysis.

Results. In Table C.2, we report the results of the Grade 6 specification applied to the Grade 10 sample under the Grade 6 sample rules: school-level score means, region-by-bilingual strata

Table C.1: Grade 10: completeness of score records by ownership and regime

	Public		Non-public	
	Internal	External	Internal	External
Students in file	12,774	8,204	6,999	3,849
Complete records (all 4 subjects, %)	72.4	69.7	92.3	93.5
No scores at all (%)	21.6	20.5	5.0	4.8
Partial records (1–3 subjects, %)	5.9	9.8	2.7	1.7
Mean Spanish score, partial records	358	201	503	303
Mean Spanish score, complete records	485	500	528	522

Notes: Randomized (non-Capital) sample; students as recorded in the administrative file. "Complete" denotes valid scores in Spanish, English, mathematics, and social/civic competences. Scores are standardized to mean 500 and standard deviation 100 at the student level. Fewer than 5 percent of partial records carry a special-situation code.

fixed effects interacted with school size, and school shares of girls and immigrant students as controls (386 randomized schools, of which 138 external, 69 of them non-public). We focus on Panel B, which tests the central prediction of our framework: that external evaluation lowers the scores of non-public schools more than those of public schools. The heterogeneity by ownership replicates: the External×Non-public interaction is negative in the composite and in three of the four subjects, and it is statistically significant for the composite. The interaction on the composite score is about −15 points, or 15 percent of a standard deviation, a little under half the size of the corresponding Grade 6 estimate. This result, however, depends on how partial records are treated: without the completeness restrictions, which keep students with scores in only some subjects out of the sample, the Spanish and English interactions turn positive and the composite interaction is close to zero. Only the mathematics interaction remains negative and statistically significant with or without these restrictions (see the notes to Table C.2).

These results are consistent with the incentive interpretation of our main findings. Although by 2017 the Grade 10 evaluation carried no academic consequences for students, its results still matter to families: Grade 10 marks the end of compulsory schooling, and a school's results at this stage speak to how well it prepares students for upper-secondary education and, eventually, for university. Non-public schools thus have reputational incentives at this grade as well, and the Grade 10 wave corroborates the ownership heterogeneity out of sample, most robustly in mathematics. Its incomplete score records, however, prevent us from establishing the comparability and balance that support the Grade 6 analysis, which is why we base the paper's headline results on the Grade 6 data.

Table C.2: Grade 10: effects of external evaluation on school-average scores

	Composite	Spanish	English	Mathematics	Soc./Civic
Panel A: average effect					
External	−1.00	3.83	6.33	−6.94	−7.24
	(3.74)	(3.71)	(4.29)	(4.62)	(5.36)
Non-public	41.74***	25.00***	57.42***	36.05***	48.49***
	(5.42)	(4.85)	(5.83)	(7.15)	(7.37)
Panel B: heterogeneity by ownership					
External	6.51	14.84**	11.72**	6.15	−6.66
	(4.68)	(5.72)	(5.17)	(4.51)	(6.00)
Non-public	46.70***	32.27***	60.98***	44.69***	48.87***
	(6.33)	(5.82)	(6.57)	(8.67)	(8.30)
External×Non-public	−14.97**	−21.94***	−10.74	−26.07***	−1.15
	(7.19)	(7.12)	(8.38)	(8.59)	(10.53)
Observations	386	386	386	386	386

Notes: School-level regressions on the randomized (non-Capital) sample under the Grade 6 sample rules: students with valid scores in all four subjects, schools with at least 30 percent participation and seven valid students, and non-missing bilingual status (386 schools; 138 external, of which 69 non-public). All specifications absorb region-by-bilingual strata fixed effects interacted with school size and include school shares of girls and immigrant students. Standard errors clustered at the school level in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. The composite is the mean of the four subject scores. All estimates below include the same controls (at the student level, indicators for girls and immigrant students). The mathematics interaction is robust to lifting the completeness restrictions (−23.6, $p = 0.006$), to student-level estimation with school-clustered standard errors (−26.8, $p = 0.001$), and to restricting the sample to students in the academic mathematics track (−26.9, $p = 0.003$). The Spanish and English interactions, and with them the composite, are sensitive to the treatment of partial records (see Table C.1): without the completeness restrictions they are +14.1 ($p = 0.21$), +9.1 ($p = 0.32$) and +0.3 ($p = 0.97$).

D Derivation of the Ranking Implications

This appendix derives the results on rankings stated in Section 4.1 and the two implications that follow from them in Section 4.2. We first describe how measured performance determines the ranking, then derive the percentile rank of each school and the composition of the top of the ranking, and finally show how monitoring changes both.

D.1 Setup

There are two school types, public (P) and non-public (N), and a share λ of schools is non-public. Quality is distributed across public schools according to G_P, with density g_P, and across non-public schools according to G_N, with density g_N. Both distributions are continuous, so two schools have the same quality with probability zero.

As in Section 4.1, all schools of type i choose the same manipulation level s_i^*, and the measured performance of a school of type i with quality q is $m = q + s_i^*$. The manipulation gap is $\Delta = s_N^* - s_P^*$, which is positive because $b_N > b_P$.

A ranking depends only on the order of measured performance, and adding the same amount to every school's score does not change that order. We can therefore subtract s_P^* from all scores and work with

$$\tilde{m} = q \quad \text{for public schools,} \qquad \tilde{m} = q + \Delta \quad \text{for non-public schools.} \tag{10}$$

Everything that follows depends on manipulation only through Δ: what matters for rankings is not how much schools manipulate, but how much more non-public schools manipulate than public ones.

D.2 Percentile ranks

The percentile rank of a school is the share of all schools with lower measured performance. Take a non-public school with quality q. Two groups of schools rank below it. Among non-public schools, those with lower quality, which are a share $G_N(q)$ of non-public schools and therefore a share $\lambda G_N(q)$ of all schools. Among public schools, those with quality below $q + \Delta$, which are a share $G_P(q + \Delta)$ of public schools and therefore a share $(1 - \lambda)G_P(q + \Delta)$ of all schools.

Adding up, the percentile rank of the non-public school is

$$R_N(q) = \lambda\, G_N(q) + (1-\lambda)\, G_P(q+\Delta). \tag{11}$$

The first term reflects comparisons with other non-public schools and does not depend on Δ. The second reflects comparisons with public schools and increases with Δ. By the same reasoning, a public school with quality q ranks above public schools with lower quality and above non-public schools with quality below $q-\Delta$, so its percentile rank is

$$R_P(q) = (1-\lambda)\, G_P(q) + \lambda\, G_N(q-\Delta). \tag{12}$$

To obtain the average percentile rank of non-public schools, we average Equation (11) over their qualities:

$$E[R \mid N] = \lambda\, E[G_N(q_N)] + (1-\lambda)\, E[G_P(q_N+\Delta)]. \tag{13}$$

The first expectation is the average rank of a non-public school among non-public schools only. Ranks within any group of schools are spread evenly between zero and one, so they average one half: formally, $G_N(q_N)$ is uniformly distributed on $[0,1]$ when G_N is continuous, and $E[G_N(q_N)] = 1/2$. The second expectation is the probability that a randomly drawn non-public school ranks above a randomly drawn public school. By Equation (10), a non-public school with quality q_N ranks above a public school with quality q_P if and only if $q_N + \Delta > q_P$, so this probability is

$$\pi(\Delta) = \Pr(q_N + \Delta > q_P) = \int G_P(q+\Delta)\, dG_N(q), \tag{14}$$

where the second equality conditions on the quality q of the non-public school, which ranks above every public school with quality below $q+\Delta$. Since G_P is increasing, so is π, with $\pi'(\Delta) = \int g_P(q+\Delta)\, dG_N(q) \geq 0$. Therefore,

$$E[R \mid N] = \frac{\lambda}{2} + (1-\lambda)\, \pi(\Delta), \tag{15}$$

which increases with Δ. The same steps applied to Equation (12), using $E[G_N(q_P-\Delta)] =$

$\Pr(q_N + \Delta < q_P) = 1 - \pi(\Delta)$, give the average percentile rank of public schools,

$$E[R \mid P] = \frac{1-\lambda}{2} + \lambda\,[1 - \pi(\Delta)], \tag{16}$$

which decreases with Δ. Since percentile ranks average one half across all schools, $\lambda E[R \mid N] + (1-\lambda)E[R \mid P] = 1/2$ for any Δ: the rank gains of non-public schools are exactly offset by the rank losses of public schools.

D.3 Composition of the top of the ranking

Consider the top x share of the ranking, for instance $x = 0.1$ for the top decile. Its cutoff c is the score above which a share x of all schools lies. By Equation (10), a non-public school is above the cutoff if its quality exceeds $c - \Delta$, and a public school if its quality exceeds c, so c solves

$$\lambda\,[1 - G_N(c - \Delta)] + (1-\lambda)\,[1 - G_P(c)] = x. \tag{17}$$

The first term is the share of schools above the cutoff that are non-public, and the second the share that are public. A larger Δ moves non-public schools up, so for a given cutoff more schools lie above it. To keep the top group at size x, the cutoff must rise. By the implicit function theorem,

$$\frac{dc}{d\Delta} = \frac{\lambda\, g_N(c - \Delta)}{\lambda\, g_N(c - \Delta) + (1-\lambda)\, g_P(c)} \in [0, 1]. \tag{18}$$

As the cutoff rises, fewer public schools remain above it, and since the size of the top group is fixed, the share of non-public schools in the group rises. Formally, the share of non-public schools in the top group is

$$S(\Delta) = 1 - \frac{(1-\lambda)\,[1 - G_P(c)]}{x}, \qquad S'(\Delta) = \frac{(1-\lambda)\, g_P(c)}{x}\,\frac{dc}{d\Delta} \geq 0. \tag{19}$$

This holds for any x. When $\Delta = 0$, S equals the share of non-public schools in the top group of the ranking by true quality.

D.4 Where rank changes concentrate

Equation (11) also shows which non-public schools gain the most from manipulation. Differentiating with respect to Δ,

$$\frac{\partial R_N(q)}{\partial \Delta} = (1-\lambda)\, g_P(q + \Delta). \tag{20}$$

A non-public school gains one percentile point for each percent of all schools it overtakes. A small increase in Δ moves it past the public schools whose quality is just above $q+\Delta$, and the number of such schools is proportional to the density $g_P(q+\Delta)$. The same increase in manipulation therefore moves a school far in the ranking when many public schools have similar measured performance, as in the middle of the distribution, and moves it little in the tails. In particular, a non-public school that already ranks near the top has few public schools left to overtake. Similarly, $\partial R_P(q)/\partial\Delta = -\lambda\, g_N(q-\Delta)$: public schools lose the most rank where non-public schools are most densely packed.

D.5 Monitoring and Implications 3 and 4

Section 4.1 shows that manipulation falls with monitoring for both types, and faster for non-public schools: $|\partial s_N^*/\partial f| > |\partial s_P^*/\partial f|$. The manipulation gap therefore falls with monitoring,

$$\frac{\partial \Delta}{\partial f} = \frac{\partial s_N^*}{\partial f} - \frac{\partial s_P^*}{\partial f} < 0, \tag{21}$$

so that $\Delta(f_{ext}) < \Delta(f_{int})$. Combined with the results above, this yields the two implications in Section 4.2. Since $E[R \mid N]$ increases with Δ, the average percentile rank of non-public schools is lower under external evaluation (Implication 3). Since the share of non-public schools in any top group increases with Δ, that share is also lower under external evaluation (Implication 4).

Two remarks help interpret these implications. First, they compare the ranking when all schools are evaluated under one regime with the ranking when all schools are evaluated under the other. Second, rankings respond only to the difference in manipulation between types. If external evaluation reduced manipulation by the same amount for public and non-public schools, scores would fall (Implication 1), but rankings would not change.

D.6 Benefits that depend on quality

Suppose that the marginal benefit of manipulation also depends on quality, $b_i(q)$, for instance because the returns to reaching the top of the ranking are larger. Manipulation then varies across schools of the same type, $s_i^*(q,f)$, and the ranking can change within type as well as across types. Let

$$r_i(q) = s_i^*(q, f_{int}) - s_i^*(q, f_{ext}) \geq 0 \tag{22}$$

denote the fall in the measured performance of a school of type i and quality q when moving from internal to external evaluation. Implications 3 and 4 continue to hold if external evaluation lowers the measured performance of every non-public school by at least as much as that of every public school, $r_N(q) \geq r_P(q')$ for all q and q'.

For Implication 3, note that Equation (15) does not rely on schools of the same type keeping their order: whatever the order among non-public schools, their ranks among themselves average one half, so $E[R \mid N] = \lambda/2 + (1-\lambda)\pi$, where π is the probability that a non-public school ranks above a public school. Under the condition above, moving to external evaluation lowers the score of each non-public school by at least as much as that of each public school. No pair of schools switches order in favor of the non-public school, so π, and with it $E[R \mid N]$, cannot rise.

For Implication 4, subtract from all scores under external evaluation a constant $\bar{r}$ between the largest $r_P(q')$ and the smallest $r_N(q)$, which does not change the ranking. Relative to internal evaluation, non-public scores are then weakly lower and public scores weakly higher. Let $N(c)$ and $P(c)$ denote the shares of all schools that are non-public and public with scores above c under internal evaluation, and $N'(c) \leq N(c)$ and $P'(c) \geq P(c)$ the same shares under external evaluation. Let c and c' be the cutoffs of the top x group in each case. If $c' \geq c$, then $N'(c') \leq N'(c) \leq N(c)$. If $c' < c$, then $P'(c') \geq P'(c) \geq P(c)$, and since the top group has size x in both cases, $N'(c') = x - P'(c') \leq x - P(c) = N(c)$. Either way, the non-public share of the top group cannot rise.

Finally, if $b_i(q)$ increases with q, the first-order condition $b_i(q) = \partial C_i(s, f)/\partial s$ with convex costs implies that $s_i^*(q, f)$ also increases with q. Higher-quality schools then manipulate more, and the effect of external evaluation on scores is larger in the upper part of the distribution, as in the quantile regressions of Section 6.

D.7 Interpretation of the rank regressions

The rank regressions in Section 6.3 compare the rank of a school under each regime when the other schools keep their assigned regime, most of them internal, rather than two full rankings. In the model, moving a single school of type i from internal to external evaluation lowers its measured performance by $s_i^*(f_{int}) - s_i^*(f_{ext})$, which is larger for non-public schools, and lowers its rank by the share of schools it falls behind. The rank regressions therefore test the same prediction as Implications 3 and 4 under the assignment actually observed. Their magnitudes

need not coincide with those of the full-ranking comparison, because the rank a school loses also depends on how many schools have similar scores.

E DFL Methodology

This appendix describes the counterfactual score distributions in Section 6.2. Section E.1 adapts the derivation of DiNardo et al. (1996) to our setting, and Section E.2 describes the estimation.

E.1 The counterfactual distribution

Each observation is a school s with three components: its average score in a given subject, y_s, a vector of characteristics, z_s, and an indicator E_s that equals one if the school was externally evaluated and zero if it was internally evaluated. All objects in this appendix are defined within an ownership type, public or non-public, and we omit the corresponding index for brevity. The characteristics z are the design variables of the experiment (Section 6.2).

We write $f(y; E_y = e, E_z = e')$ for the density of scores generated under regime e when the distribution of characteristics is that of schools evaluated under regime e', where $e, e' \in \{0, 1\}$. When $e = e'$, this is the actual density of scores among schools evaluated under regime e, and by the Law of Iterated Expectations,

$$f(y; E_y = e, E_z = e) = \int_z f(y \mid z, E_y = e)\, dF(z \mid E_z = e), \tag{23}$$

where $f(y \mid z, E_y = e)$ is the density of scores conditional on characteristics under regime e, and $F(z \mid E_z = e)$ is the distribution of characteristics among schools evaluated under regime e. The actual densities under internal and external evaluation are therefore

$$f_{\text{int}}(y) \equiv f(y; E_y = 0, E_z = 0), \qquad f_{\text{ext}}(y) \equiv f(y; E_y = 1, E_z = 1). \tag{24}$$

The counterfactual density, $f_{\text{cf}}(y)$, is the density of scores under external evaluation had the distribution of characteristics been that of internally evaluated schools:

$$\begin{aligned} f_{\text{cf}}(y) \equiv f(y; E_y = 1, E_z = 0) &= \int_z f(y \mid z, E_y = 1)\, dF(z \mid E_z = 0) \\ &= \int_z f(y \mid z, E_y = 1)\, \Psi(z)\, dF(z \mid E_z = 1), \end{aligned} \tag{25}$$

where we have made use of a reweighting function defined as

$$\Psi(z) = \frac{dF(z \mid E_z = 0)}{dF(z \mid E_z = 1)}. \tag{26}$$

The counterfactual $f_{\text{cf}}(y)$ is the distribution of scores that internally evaluated schools would display under external grading. It is valid if the conditional density of scores under external evaluation, $f(y \mid z, E_y = 1)$, is the same for internally evaluated schools as for the externally evaluated schools from which it is estimated. In DiNardo et al. (1996) this is an assumption. Here it holds by design, because the regime was randomly assigned within strata defined by z.

Because estimating Expression (26) directly would be imprecise, given the small number of schools with each value of z, we apply Bayes' rule. For $e \in \{0, 1\}$,

$$dF(z \mid E_z = e) = \frac{\Pr(E = e \mid z)\, dF(z)}{\Pr(E = e)}. \tag{27}$$

Writing $p(z) \equiv \Pr(E = 1 \mid z)$ for the probability of external evaluation given z, the reweighting function reads

$$\Psi(z) = \frac{\Pr(E = 0 \mid z)\ \Pr(E = 1)}{\Pr(E = 1 \mid z)\ \Pr(E = 0)} = \frac{1 - p(z)}{p(z)} \cdot \frac{\Pr(E = 1)}{\Pr(E = 0)}, \tag{28}$$

so that each externally evaluated school is weighted by its odds of internal evaluation. The counterfactual therefore requires only a model for $p(z)$.

Finally, we compare the counterfactual with the actual density of scores under internal evaluation:

$$\Delta(y) = f_{\text{int}}(y) - f_{\text{cf}}(y) = f(y; E_y = 0, E_z = 0) - f(y; E_y = 1, E_z = 0). \tag{29}$$

Both densities use the distribution of characteristics of internally evaluated schools and differ only in the regime under which scores are generated. $\Delta(y)$ is the difference plotted in Figures 2 and A.5.

E.2 Estimation

Steps. We use the sample of the main analysis. For each outcome (the average score in each subject and the composite score) and each ownership type, we proceed in four steps:

1. **Propensity score.** We estimate $p(z)$ by a logit of E_s on the district-by-bilingual strata indicators, the number of students in grade 6, and its interactions with the strata indicators.

2. **Weights.** Each externally evaluated school receives the weight $\hat{\Psi}_s = (1 - \hat{p}(z_s))/\hat{p}(z_s)$. The factor $\Pr(E = 1)/\Pr(E = 0)$ in Equation (28) is common to all schools and drops out once the weights are normalized, $\hat{\omega}_s = \hat{\Psi}_s / \sum_{r:E_r=1} \hat{\Psi}_r$.

3. **Densities.** We estimate the actual densities by kernel density estimation on internally and externally evaluated schools, respectively, and the counterfactual density by the weighted kernel estimator on externally evaluated schools:

$$\hat{f}_{\text{int}}(y) = \frac{1}{N_0} \sum_{s:E_s=0} K_h(y - y_s),$$
$$\hat{f}_{\text{ext}}(y) = \frac{1}{N_1} \sum_{s:E_s=1} K_h(y - y_s),$$
$$\hat{f}_{\text{cf}}(y) = \sum_{s:E_s=1} \hat{\omega}_s \, K_h(y - y_s),$$

   where N_0 and N_1 are the numbers of internally and externally evaluated schools, $K_h(u) = K(u/h)/h$, and K is the Epanechnikov kernel.

4. **Difference.** We compute $\hat{\Delta}(y) = \hat{f}_{\text{int}}(y) - \hat{f}_{\text{cf}}(y)$ on a grid of 300 equally spaced points spanning the range of observed scores, extended by 6% on each side.

Bandwidth. All densities of a given outcome use the same bandwidth, $h = 1.06\,\hat{\sigma} N^{-1/5}$, where $\hat{\sigma}$ is the standard deviation of the outcome and N the number of schools in the pooled sample of both regimes and ownership types. A group-specific bandwidth would adapt the degree of smoothing to each group, but differences between curves would then partly reflect different smoothing. With a common bandwidth, differences between the three densities, and between public and non-public schools, reflect the data alone. The cost is that the densities of the small groups, in particular the 22 externally evaluated non-public schools, are somewhat noisier than a group-specific rule would make them.[20]

Choice of covariates. The counterfactual density depends on which covariates enter the propensity score, and the choice involves a trade-off. Conditioning on more school characteristics adjusts for chance imbalances in them, but in a small sample it also pushes the estimated propensity score of some externally evaluated schools close to zero. Because each of these schools receives the weight $(1-\hat{p})/\hat{p}$, a few schools then carry a large share of the weight, and the coun-

[20] Our implementation reproduces the output of the Stata command `dfl` when both use the same kernel and bandwidth.

terfactual density rests mostly on them. We measure this cost with two diagnostics. The largest normalized weight shows whether a single school dominates the counterfactual. The effective number of schools, $(\sum_s \hat{\Psi}_s)^2 / \sum_s \hat{\Psi}_s^2$, equals the number of externally evaluated schools when all weights are equal and falls as the weight concentrates on fewer schools. In our setting, the trade-off favors restricting the propensity score to the design variables. Because the regime was randomly assigned within strata, these are the only variables on which the probability of external evaluation depends, and conditioning on them is enough for the counterfactual in Equation (25) to be valid. Adding further school characteristics would only adjust for chance imbalances in them, and with 22 externally evaluated non-public schools it overfits the propensity score. Table E.1 compares the design-only specification with one that also includes the school averages of the seven student characteristics in Table 2. With the additional controls, the largest normalized weight among non-public schools is 7 times the mean, and the effective number of externally evaluated non-public schools falls to 7.7 of 22. With the design variables only, no non-public school receives more than 1.7 times the mean weight, and the effective number is 18.5 of 22. For public schools, the two specifications give similar weights.

Table E.1: Propensity Score Diagnostics for the DFL Counterfactual

		Range of $\hat{p}$		Largest	Effective number of
Propensity covariates	Schools	External	Internal	normalized weight	external schools
Design variables	Public	[0.117, 0.408]	[0.082, 0.377]	1.78	54.7 of 58
(baseline)	Non-public	[0.143, 0.811]	[0.012, 0.425]	1.67	18.5 of 22
Design variables +	Public	[0.083, 0.441]	[0.064, 0.496]	2.58	48.1 of 58
school composition	Non-public	[0.037, 0.837]	[0.000, 0.494]	6.99	7.7 of 22

Notes: Logit of external evaluation estimated separately by ownership type on the 408 schools of the main sample. Design variables: district × bilingual strata, number of grade-6 students, and their interactions. School composition: school averages of the seven student characteristics in Table 2. Weights $\hat{\Psi}_s = (1 - \hat{p})/\hat{p}$ for externally evaluated schools, normalized to mean one. Effective number of schools: $(\sum_s \hat{\Psi}_s)^2 / \sum_s \hat{\Psi}_s^2$.