Martins, Pedro S.; Ferreira, João R.

Working Paper

Effects of Individual Incentive Reforms in the Public Sector: The Case of Teachers

GLO Discussion Paper, No. 1441

Provided in Cooperation with:
Global Labor Organization (GLO)

Suggested Citation: Martins, Pedro S.; Ferreira, João R. (2024) : Effects of Individual Incentive Reforms in the Public Sector: The Case of Teachers, GLO Discussion Paper, No. 1441, Global Labor Organization (GLO), Essen

This Version is available at:
https://hdl.handle.net/10419/296494

Effects of Individual Incentive Reforms in the Public Sector: The Case of Teachers*

Pedro S. Martins†
Nova School of Business and Economics
& IZA & GLO

João R. Ferreira‡
Nova School of Business and Economics

May 30, 2024

Abstract

We evaluate a political reform in Portugal that introduced individual teacher performance-related pay and tournaments in public schools. We find that the focus on individual performance decreased student achievement, as measured in national exams, and increased grade inflation. The results follow from a difference-in-differences analysis of matched student-school panels and two complementary control groups: public schools in regions that were exposed to lighter reforms; and private schools, whose teachers had their incentives unchanged. Students in public schools with a higher proportion of teachers exposed to the tournament also perform worse. Overall, our results highlight the potential social costs from disruption of cooperation amongst public sector workers due to competition for promotions.

Keywords: Tournaments, Public Sector, Teacher Merit Pay, Matched School-Student Data.

JEL Codes: I21, M52, I28.

*We thank Pedro Pita Barros, Nuno Crato, Michaela Gulemetova, Victor Lavy, Helena Martins, Álvaro A. Novo, Phil Oreopoulos, Thomas Piketty, Pedro Portugal, Rodrigo Queiros e Melo, Ron Zimmer, and seminar participants at the Institute of Education (London), ISEG (Lisbon), and Nova SBE (Carcavelos) for helpful comments and discussions. We thank support from Fundação para a Ciência e a Tecnologia (UIDB/00124/2020, UIDP/00124/2020 and Social Sciences DataLab - PINFRA/22209/2016), POR Lisboa and POR Norte (Social Sciences DataLab, PINFRA/22209/2016). Any errors are our own.

†Corresponding author. Email: pedro.martins@novasbe.pt. Address: Nova School of Business and Economics, Universidade NOVA de Lisboa, Campus de Carcavelos, 2775-405 Carcavelos, Portugal. Web: https://pmrsmartins.wixsite.com/website.

‡Email: joao.ferreira@novasbe.pt. Address: Nova School of Business and Economics, Universidade NOVA de Lisboa, Campus de Carcavelos, 2775-405 Carcavelos, Portugal.

1 Introduction

As education is a key area of government intervention in most countries, the design of incentives for public-sector teachers is a major challenge for governments (Lazear 2003). While several studies confirm that teachers can make a significant difference in students' achievement (Rivkin et al. 2005, Chetty et al. 2014), it has proved difficult to understand the drivers of the differences in teacher quality (Aaronson et al. 2007), including the best format of their incentives. This paper provides novel evidence on how public policies (and politics) around teacher incentives can shape students' achievement. Our results may also inform other potential incentive reforms around the functioning of government agencies (Dixit 2002, Burgess et al. 2017).

In the education sector, teacher incentives, either individual or collective, may improve student achievement if they align public goals with teacher goals. In this case, a combination of incentive and composition effects will increase student performance (Lazear 2000, 2003). However, an approach in which reward is based on outputs or outcomes can also be fraught with difficulties, which may explain the popularity of simpler, input-based rewards (Kane & Staiger 2002). For instance, setting specific measurable outputs may lead to 'teaching to the test', which can involve dysfunctional behaviour. Moreover, while individual incentives may disrupt collaborative work (Fehr & Schmidt 1999), collective incentives may also generate free riding and, in the end, little effect on performance.

In addition to the theoretical ambiguity of the effects of teacher incentives, the extant empirical literature on this topic has not come to definitive conclusions (see Pham et al. (2021) for a meta-analysis of research on teacher merit pay). Lavy (2002, 2009), Dee & Wyckoff (2015) are early references that explore randomized or quasi-natural experiments to address the causal relationship between teacher incentives and student achievement, providing support for the potential of collective and individual incentives. Sojourner et al. (2014) finds positive effects from pay-for-performance schemes in Minnesota, especially amongst less experienced teachers. Eren (2019) evaluates a hybrid programme in Louisiana, involving both group and individual incentives, finding significant effects on achievement in Maths. On the other hand, Fryer (2013), Goodman & Turner (2013), Imberman & Lovenheim (2015), and Leone (2024) present analyses of school-level or group-based teacher incentive systems, finding non-significant or (at most) modest effects of such compensation schemes, particularly when incentives to free

ride are stronger and when students are assessed in low-stakes exams.[1] Recent research has focused on improving specific design aspects of incentives programmes. In a behavioural field experiment in Illinois, Fryer et al. (2022) find that financial incentives for teachers may indeed be effective if their design induces loss aversion. Several other studies focus on the cases of developing countries, also with mixed findings.[2]

While the literature above generally focuses on regional pilot projects, here we examine the effects of the introduction of individual teacher incentives in all public-sector schools in a given country, Portugal. The main aspects of this reform, described in more detail in Section 2, are the breaking up of the until then single pay scale for teachers into two pay scales and the tournament-like structure for progression between the lower and upper pay scales. While before the reform progression (and wage growth) depended almost only on tenure, the new incentives placed considerable emphasis on the school-level and national-exam results of the students taught by each teacher. However, as any progressions between the two pay scales could not exceed a given number of upper-scale teacher vacancies per school determined centrally by the Ministry of Education, the incentive structure amounted to a form of tournament (Lazear & Rosen 1981). Overall, these changes established a clear contrast in the incentives faced by public-school teachers, from inputs (hours and years of work) to individual outputs (the grades of the students of each teacher).

We study the effects of this reform on students' school-level and national-exam results. Specifically, we draw on matched student-school data covering the population of secondary school students that sat national exams from 2002 to 2011. We then conduct a difference-in-differences (DID) analysis based on two complementary control groups. In the first control

[1] See Jones (2013) for an analysis that considers additional outcomes such as teacher work hours and turnover. See also Figlio & Kenny (2007) for an early study of the US based on cross-sectional data and Atkinson et al. (2009) for an analysis of the introduction of performance-related pay for teachers in England. See Brehm et al. (2017) for an analysis of an individual merit pay tournament in Houston, where performance was measured by teacher value added. Bergman & Hill (2018) considers a related approach of making teachers' value-added ratings available to the public. See also Barlevy & Neal (2012) for a theoretical analysis, focusing on collective teacher incentives.

[2] Loyalka et al. (2019), for instance, compare 'pay-for-percentile' incentives ("which reward teachers based on the rankings of individual students within appropriately defined comparison sets," p. 623) with more frequent approaches, such as those focusing on average class performance or average improvement in scores throughout a school year. The former design led to better results. A randomized experiment in Kenya (Glewwe et al. 2010) is not supportive of the role of teacher incentives while a similar study of India (Muralidharan & Sundararaman 2011) is. Behrman et al. (2015) finds positive achievement effects in Mexico schools but mostly when both individual and group incentives are provided to different stakeholders, including students, teachers, and school administrators. Barrera-Osorio & Raju (2017) does not find achievement effects from a randomized controlled trial of a pilot teacher performance pay programme in Pakistan. In a recent randomized evaluation in Tanzania, Mbiti et al. (2019) show that a combination of unconditional grants (i.e., increased spending on school inputs) and pecuniary incentives based on student performance (for teachers) can be more effective than separate interventions.

group, we consider public schools in two insular autonomous regions (Azores and Madeira). These regions were exposed to lighter versions of the reform than the rest of the country, as their pay scale remained unchanged and teacher progression was less restricted. In the second control group, we consider private schools. The students of these schools are subject to the same national exams as the treatment group but their teachers were not affected by the reform: pay and any incentives remained freely set by each private school, subject only to wage floors determined by collective bargaining.

Our research contributes to the literature on the effects of public sector and teacher incentives in different ways. First, this is one of the few studies of a full reform (rather than a pilot project) in a developed economy and that conducts the analysis drawing on population data. We are therefore able to address issues of external validity that arise in experimental settings[3] and that typically receive little attention due to data constraints, particularly in the empirical incentives literature (Lazear 2000, Bandiera et al. 2005). Indeed, our population data contrasts with the studies mentioned above that draw on randomized or quasi-experimental studies (Lavy 2002, 2009, Glewwe et al. 2010, Muralidharan & Sundararaman 2011, Dee & Wyckoff 2015). Moreover, our analysis of systematically-collected official data throughout the period will alleviate measurement error bias and Hawthorne effects compared to the case of a typical experiment.

A second important aspect of our study involves the several robustness tests, on top of the two complementary control groups. For instance, our analysis of a long period of up to five years before the reform allows us to test the common trends hypothesis in some detail. Furthermore, we consider a number of different specifications that control for different sets of variables, including school and school-exam fixed effects, and different data subsets. We also leverage the administrative student- and teacher-school matched panel dataset to discuss possible compositional effects, namely on the student bodies of public and private schools. Moreover, we compare public schools depending on their intensity of competition for promotions to the upper pay scale, given the (more or less advanced) career stage of their tenured teachers.

Finally, we pay particular attention to the potential for grade inflation (Jacob & Levitt 2003), measured by the gaps between external and internal grades of the same students. Grade

[3] "[A] weakness of natural experiments is that their results may not be generalizable beyond the group of individuals or firms or the setting used in the study" (Meyer 1995).

inflation may emerge from the fact that the progression criteria are affected by student results that are determined in part by the teachers themselves.

Overall, we find that the increased focus on individual teacher performance caused a statistically significant decline in student achievement. This decline in achievement is more pronounced in the case of national exams, with an effect of about one fifth of a standard deviation. Consistently with the different effects in terms of internal and external results, our triple-difference evidence documents a significant increase of grade inflation. In addition, we find that there are no significant differences between the treatment and control groups' trends before the introduction of merit pay. Moreover, the negative effects upon national exams are cumulative in the period after the reform is introduced. Leveraging administrative data, we find that compositional changes in the exam-taking population (with respect to public and private schools) do not seem to be a concern. We also present evidence that the negative effects of the reform were more pronounced in schools were competition for promotions was more intense. This is consistent with a decrease in cooperation among teachers, which is a plausible theoretical mechanism behind the relative decline in public-school students' achievement after the reform. Finally, the inclusion of different control variables or the consideration of different subsets of the data makes only very minor differences to the size of our estimates, as would be the case if assignment to treatment were random.[4]

The structure of the paper is as follows: Section 2 describes the main characteristics of the education reform studied in the paper and discusses some of its theoretical implications. Section 3 presents the data used in the paper, a matched school-student panel data set; Section 4 describes the main results, while Section 5 presents our robustness analyses. Finally, Section 6 concludes.

2 The teacher incentives reform

A new government that came into office in 2005 decided to respond to the evidence of relatively poor performance levels in the Portuguese education system (OECD 2001). Indeed, when evaluated in international assessments such as the PISA tests, students in Portugal did not

[4]These robustness analyses add to the earlier finding that teachers in those public schools that exhibited larger falls in performance after the reform were more likely to take (costly) early retirement when it became available in 2008 (Martins 2010). This is consistent with the potential negative effects of the incentives/merit pay scheme on cooperation amongst teachers, administrative workload and, in the end, overall job satisfaction, as suggested by theory.

fare well. This was particularly concerning when taking into account the relatively high public expenditure levels in education, of which relatively high average teacher salaries were an important component.

A key aspect of the education reform introduced by that government was the breaking up of the single pay scale for teachers into two separate scales. This and other aspects of the reform became law in January 2007, after having been subject to public discussion for several months and approved by the government in November 2006. The breakup of the pay scale marked an important contrast with the period before the 2006/07 school year as teachers were no longer ensured of virtually automatic, tenure-related progression from the bottom to the top of the pay scale over their careers. In particular, the gap between the last point in the lower scale and the first point of the higher scale was particularly large, at around 25%, from about €2,000 to about €2,500 per month (gross). Those teachers in the higher pay scale were supposed to play a special role in management and pedagogical tasks in their schools.

Another key aspect of the reform is that the new system conditioned progression from the lower to the upper pay scale on a number of *individual* teacher performance variables. These criteria were virtually nonexistent until them. One such criterion for teacher progression, which received by far most media attention, was the academic performance of the students taught by each teacher. Another criterion that also received considerable attention was the feedback from the students' parents about the teacher. The remaining criteria included the teacher's attendance record, attendance at training sessions, management and pedagogical duties, and involvement in research projects.

According to the new law, these criteria for progression were to be assessed at each school, by those teachers in the higher pay scale. Moreover, detailed assessment sheets were made available by the Ministry of Education to be used when gathering information on the above-mentioned criteria. However, even if the teacher did well along these criteria, progression between the two pay scales was still conditional on a given number of (upper-scale) teacher vacancies per school, determined centrally every two years by the Ministry of Education as a function of the number of students in the school. On the political economy side, the reform generated heated debate and opposition from teacher unions and many teachers, including two national strikes.

From the above, we conclude that the reform under study involves a stark contrast in

teacher assessment and incentives. In particular, the new framework introduced several aspects which can be characterised not only as (individual) performance-related pay but also more specifically as tournaments (Lazear & Rosen 1981). Doing extremely well may not be enough for a promotion if one's colleagues do even better and take all promotion opportunities available. Turning to a theoretical discussion of the predicted effects of this reform, there are different arguments to take into account. First, the greater weight placed on performance indicators would presumably induce teachers to focus their effort on those criteria highlighted in the law. This is expected to increase student achievement, which is measured by national exams and, to a lesser extent, school-level results.

On the other hand, tournaments are known to be potentially disruptive in terms of the collaborative work amongst agents involved in a competition (see Martins (2008) and the references therein). Moreover, collaborative work may be particularly important in the public sector, especially in education. Fairness concerns may come to the fore and undermine teacher morale (Fehr & Schmidt 1999, 2004), given the difficulties in assessing teacher contributions (Jacob & Lefgren 2008). Moreover, setting broadly measurable outputs may lead to dysfunctional behaviour such as grade inflation, particularly in internal (school-level) marks, which are determined by teachers.[5] Finally, the administrative burden involved in this or any other teacher assessment process may also be considerable. For instance, the time spent handling the formal application for progression including the required supporting documents may reduce the effort that teachers put into teaching activities.[6]

One important additional aspect of the education reform studied here is that it applied only to a smaller extent in the two autonomous regions in Portugal, the Azores and Madeira islands. Indeed, these two regions have legislative powers in education, and they decided not to follow the education reform in the mainland. Specifically, the two regions also introduced greater emphasis on teacher assessment, under broadly the same criteria as in the mainland, although less so in the case of Madeira. However, one important difference that applied to both Azores and Madeira is that they did not break up their pay scale. These differences in

[5]Tournaments will also generate extra risk in pay which would need to be compensated by higher wages in competitive markets, but not necessarily in the regulated, public-sector labour market we study here.

[6]Indeed, teachers and other stakeholders complained frequently about this aspect of the reform. For instance, a national parents' association expressed publicly its concern about the negative effects of the reform in terms of student learning, as observed by their members. See *'Teachers' evaluation compromises students' learning, parents say'*, in newspaper *Público*, 7 Nov 2008. There were also hundreds of internet blog entries written by aggrieved teachers complaining about the increased administrative workload and diminished collegiality in their schools.

the intensity of the treatment are exploited in our empirical analysis.[7]

Furthermore, the reform did not apply at all to private schools, which account for almost one fifth of all secondary schools in the country. Teachers in private schools are rewarded independently according to the practices adopted by each school, following wage floors set by collective bargaining between private-sector school employer associations and national teacher unions.[8] In particular, we could not find any evidence of systematic changes to the personnel policies of private schools over the period or of any effect from the new teacher incentives in public schools upon the functioning of private schools, although it is difficult to rule out completely potential spillovers from mobility of teachers across school types.

3 Data

Our data cover the population of high-school national exams in Portugal over ten school years, from 2001/02 to 2010/11. (We exclude national exams data from 2012 because a public sector pay (and promotions) freeze was in effect, thereby suspending individual teacher incentives.) The data are made available by the National Exams Committee (JNE, *Júri Nacional de Exames*), an agency of the Ministry of Education which is responsible for all matters regarding the national exams carried out in the country. Upper secondary school national exams, studied in this paper, were then required for the award of the high-school diploma (for students in the academic track) and university entry (European Commission 2007).

The data include information about the internal grades obtained by students in each module (a specific subject of study, such as Portuguese or Maths) from their schools, which are based on test scores and other criteria adopted by each teacher. There is also information about students' final results in each module, after taking into account each student's internal and national-exam grades (with weights of 70% and 30%, respectively). Internal grades are truncated below the passing threshold, 10 (in a scale of 0 to 20), in which case the student cannot sit the national exam, except in special circumstances.[9]

[7] The relevant legal documents are Laws (*Decretos-Lei*) 17/2007, of January 19th and 200/2007, of May 22nd, and Regional Laws (*Decretos Legislativos Regionais*) 28/2006/A, of August 8th; 21/2007/A, of August 30th; and 6/2008/M, of February 25th. The significant distance between these two regions and the mainland (about 1600km and 950km, respectively) also minimises any possible spillover effects from the treated to the control groups (e.g., teacher mobility).

[8] Only about one fourth of these private schools are religious. See Neal (1997) for an analysis of these schools in the US context and Martins (2023) for an analysis of pay in private schools in Portugal.

[9] All data used in the paper were at some point freely available from the Education De-

Each observation concerns a unique student-module-school-year combination. Typically, there will be several observations for each student but it will not be possible to match them as the data do not include any individual student identifier. However, all schools and all modules are identified by name and unique time-invariant codes. Importantly, there is also information on the school's location at the *concelho* level (308 geographical areas) and the school's public or private status.[10]

We create our main sample of analysis by drawing on all student-exam pairs that meet the following four conditions: a first sit in the first call of a student that is applying to university and is also enrolled in the module of the exam in the school where they are sitting the exam. These criteria are similar to those adopted by most media when compiling school rankings. Our criteria are also imposed in order to ensure that the effect in terms of internal and external grades are based on the same sample and thus are strictly comparable. The resulting 2,025,402 observations are distributed across 682 schools, of which 504 are public schools in the mainland and therefore subject to the reforms described in the previous section.[11]

Table 1 presents descriptive statistics based on school (top panel) or student-exam (bottom panel) data. Amongst other results, we find that the mean internal grade is larger than 13 while the mean external exam result is lower than 11, both at the school- and student-module-level, in a scale of 0 to 20. This leads to an average gap between the two marks of more than 2, which is suggestive of considerable grade inflation or simply of different standards between school and external national assessment. We also find that, on average, there are 332 exams per year per school.

partment/JNE website; currently, only datasets from 2008 onwards are published online. Link: `https://www.dge.mec.pt/relatoriosestatisticas-0` (in Portuguese; last accessed 10 May 2024). The data were originally released openly so that the media could compile school rankings. For other analyses using these datasets, see Nunes et al. (2015) and Pereira dos Santos et al. (2021), for instance.

[10]There are several variables for each student-module-school-year combination: if the exam is a resit (either because the student failed before or because the student wants to improve their grade), if the student is applying for admission to university, and if the student is sitting the exam but is not enrolled in the school. The data also include the student's gender and age, but only for the last six years (2005/06-2010/11); and the student's school year when taking the module (typically 12^{th}, which is the last of secondary education in Portugal, but also the 11^{th}, as some modules are subject to national exams at that stage).

[11]The original size of the data is 4,242,233. 28.93% of these observations refer to second calls; 31.68% are not enrolled in the school; 7.47% are not applying for university admission; and 27.64% are resitting the exam. Of course, these exclusion categories overlap for many observations. (Extensive robustness analysis was conducted and the results presented below in Section 4 are not sensitive to different sample definitions as discussed in Section 5.) No school switches between public and private status. High-school exams (fulfilling the four conditions set out above) were sat in 140 private schools in mainland Portugal and in 36 public schools in the Azores and Madeira. 562 schools (468 public, of which 439 in the mainland and 29 in the insular regions, and 93 private in continental Portugal), comprising up to 97% of the student-module-school-year observations in the preferred analysis sample, were always present in the data throughout the ten school years we considered. We restrict the analysis sample to these "always present" schools when estimating models with school or school-exam fixed effects; the inclusion of the other schools does not qualitatively affect our results.

About 12% of the exams pertain to private schools and about 5% are from schools in the Azores and Madeira regions. Moreover, there is a downward trend in the number of exams in the period covered, which is consistent with the declining number of students enrolled in secondary school as indicated by national statistics.An exception to the trend is 2006, when new exams were introduced while some of the older exams were still sat by students.

Given that our DID estimates rely on variation over time across different groups of students, we present in Figure 1 the mean internal and external grades in each year from 2002 to 2011 at the three groups of schools we consider in our analysis: public schools in continental Portugal, public schools in the Azores and Madeira, and private schools (in the continent). We find that internal grades are very stable over the period in public schools (either in the continent or in the islands), while private schools exhibit an upward trend in the second half of the period covered. On the other hand, external marks are not only considerably lower, as documented before, but also exhibit greater fluctuation over time, including a pronounced increase across the three groups of schools from 2007 to 2008.[12] However, the increase in external marks is more pronounced in the cases of private and public/islands schools. In particular, it can be seen in Figure 1 that while the gap between internal and external marks was higher for private schools than for continent public schools in 2002-2006, this is reversed by 2007.

For additional information, Figure A1 presents the distributions of internal and external grades, focusing on the mainland and islands subsets, in 2005 (two years before the reform) and 2009 (two years after the reform). Results for other years are similar. We find that these distributions do not change in a pronounced way over the period, except perhaps for some evidence of relatively fewer very low pass internal marks. Moreover the distributions for private schools (not reported) are again very similar except that internal grades tend to follow a more uniform distribution in those schools.

[12]Starting around 2004/05, the rapid roll-out of vocational education (VET) courses in public upper-secondary schools may have changed the composition of the academic track (i.e., the group of students who have to sit national exams in order to graduate). We believe that this potential problem is mitigated for two reasons: first, the 'islands' control group would have also been affected by this change, and yet the results we obtain in that analysis are broadly in line with those for the 'private' comparison group; second, a significant proportion of early VET students were at a high risk of dropping out of school in the absence of a nonacademic track – in other words, the introduction of public-school VET did not induce a one-to-one displacement effect from the academic track (Ferreira & Martins 2023).

4 Methodology and results

4.1 Islands control group

We estimate the effects of the introduction of performance-related pay from DID models of student grade equations. Our identification assumption is that there is no effect upon grades specific to (continent) public schools with respect to the control group, from the 2006/07 school year onwards, other than from the education reform. Specifically, in the case of our first control group (Azores and Madeira), we estimate equations as follows:

$$y_{ijt} = \beta_0 + \beta_1 Continent_j + \beta_2 After_t + \beta_3 Continent_j \times After_t + u_{ijt}. \quad (1)$$

Depending on the specification, y_{ijt} denotes the (internal or external) grade of the student-exam pair i in school j in year t. Alternatively, the dependent variable is a measure of grade inflation, namely the difference between the internal and the external grade of the same student-exam pair (a triple-differences specification). Our analysis of grade inflation also serves a useful robustness purpose. Indeed, if there are other relevant interaction effects that break down the identification assumption, our results are less likely to hold across different dependent variables. More importantly, the triple-difference specification is based on a weaker identifying assumption: it simply requires that there be no shocks that affect the relative outcomes of the treatment group in the same years as the education reform.

In all cases, $Continent_j$ is a dummy variable with value one if school j is located in mainland, continental Portugal (henceforth the *continent*): this variable will pick up permanent differences in the dependent variable between schools located in the continent or in the Azores or Madeira. $After_t$ is another dummy variable, with value one if year t is 2007 (i.e., school year 2006-2007) or later, the period when the incentives reforms was in force, as discussed in Section 2: this variable will pick up across-the-board differences between the period before the intervention and the period after the intervention. This is important particularly in the case of national exam grades, as testing standards may have varied over time, as indicated in Pereira dos Santos et al. (2021).

Finally, $Continent_j \times After_t$ is the product of the two previous dummy variables; its parameter, β_3, is the object of interest in this paper. Its estimate will pick up the effect of the education reforms on student achievement or grade inflation, i.e., any additional difference

between the two types of schools that emerges after the intervention.

From the benchmark specification in equation 1, we consider three extended versions with different additional controls. The first version includes controls for school size (the total number of exams sat in each year). The second specification includes school size and school fixed effects. Finally, the third additional specification includes school-exam fixed effects. Because the structure of exams changes over the period, we focus on five topics covered in all years in these national tests: Portuguese, Maths, History, Biology & Geology, and Physics & Chemistry. Importantly, all models are estimated with robust standard errors, allowing for clustering at the school level.[13]

The first set of results, based on internal grades, are presented in Table 2 (top panel). These results draw on the student-level data described in Table 1, except that private schools are dropped. Across all four specifications, we find negative estimates for β_3, indicating that the levels of achievement of public schools in the continent fell with respect to public schools in the Azores and Madeira after the introduction of the incentives reform. The magnitude of the estimates is similar across the specifications, but small and statistically insignificant, ranging from -.009 to -.112. These values compare with a standard deviation of the dependent variable of about 2.6 marks (in a scale of 0 to 20).

However, when we turn to Table 3 (top panel), which presents similar specifications but considering external grades instead, we find much higher effects, ranging from -.296 to -.702. These correspond to as much as one fourth of a standard deviation of the dependent variable (or one half of the standard deviation of the average exam score across schools). In addition, all estimates here are also significant at the 5% level.[14] The comparison of the two sets of results (internal and external grades) therefore indicates that while national-exam results of public schools in the continent fall significantly and by a meaningful size with respect to the same change for public schools in the Azores and Madeira, the equivalent effect for internal marks is insignificant.[15]

The contrast between the internal and external results suggests that grade inflation in

[13]We exclude data from the 2006 'transition' school year from these estimations. In later year-by-year analyses, we show that our conclusions are robust to the inclusion of the 2006 cohort in our sample.

[14]Results are qualitatively similar if the dependent variable is standardized (Table B1, top panel) and when the analysis sample is expanded to include student-module-year observations that do not fulfill the four conditions set out in Section 3 (Table B2, top panel). However, in both cases, only the estimates from specifications that account for school or school-exam fixed effects remain statistically significant at the usual levels.

[15]We also find, both in Table 2 and in Table 3, that the isolated $After$ coefficients are always significantly positive, suggesting a trend towards higher marks, particularly in national exams.

the mainland's public schools is an unintended consequence of the reform. Indeed, our triple-difference estimates – see Table 4 (top panel) – indicate that the average gap between internal and external marks increases by .299 to .599 (or about one third of a standard deviation of the average gap) in public schools in the continent with respect to their counterparts in the Azores and Madeira. In all cases the coefficients are significant at the 1% level.

Even though our results are from DID models, which do not require that treated and control units be strictly comparable *ex ante*, we note that national exam scores were substantially lower in the insular regions than in the mainland before the reform. In fact, public and private school students in the mainland had much closer national-exam results. Nevertheless, the islands control group is useful for two main reasons. First, the geographical distance between the mainland and the islands reduces the likelihood of spillover effects or compositional changes following the reform that could bias the results. Second, unobserved socio-economic characteristics of students and their families that drive selection into public or private schools are not a concern in this comparison, as all students attended public schools.

Finally, it is important to recall that the Azores and Madeira also implemented lighter versions of the reform at the time. Therefore, since the islands control group was at least partially treated, we could expect the estimated effects to be downward biased – and lower than those obtained when taking the evolution of private school students as the counterfactual.

4.2 Private schools

We now turn to our second and complementary control group. Although private schools tend to exhibit better results on average in the academic achievement of their students when compared to public schools (because of a selection process, better academic practices, a greater emphasis on exam preparation, or some combination of the three), our DID approach will control for permanent differences in achievement between the two types of schools.

Similarly to the case of the first control group (equation 1), here we estimate the following DID specification:

$$y_{ijt} = \beta_0^{'} + \beta_1^{'} Public_j + \beta_2^{'} After_t + \beta_3^{'} Public_j \times After_t + u_{ijt}, \tag{2}$$

All variables take the same interpretation as before; $Public_j$ is a dummy variable with value one if school j is a State school; and $\beta_3^{'}$ is now the parameter of interest. All models in

this section are estimated with the full set of student-level data described in Table 1, except that schools located in the Azores and Madeira islands (public or private) are dropped.

Table 2 (bottom panel) presents the results for internal grades. As in the case when the public schools in Azores and Madeira served as the control group, we find again evidence that the introduction of individual teacher incentives had a detrimental effect on student achievement. However, in the present case, the coefficients display a substantially larger magnitude than in the equivalent specifications under the first control group – they range from -.411 to -.487 – and are always statistically significant, even at the 1% level.

In terms of external grades or exam results, we find that achievement in public schools, when compared to private schools, falls by between -.705 and -.889 marks after the reform. These estimates are again always significant at the 1% level – see Table 3 (bottom panel). The magnitude of these effects corresponds to about one fourth of a standard deviation of external results. The increase in the magnitude of these coefficients when compared to the results based on public schools in the islands is consistent with the intermediate intensity level of the treatment there, as discussed in Section 2 and Subsection 4.1. Results are robust to the standardization of the dependent variable (Table B1, bottom panel) and also quantitatively similar when we follow a less restrictive definition of the analysis sample (Table B2, bottom panel). In these robustness analyses, all estimates are again significant at the 1% level.

Finally, the importance of grade inflation (suggested by the stronger effects on external grades when compared to internal grades) is once again corroborated by our triple-difference results. In Table 4 (bottom panel), we find that grade inflation increases by between .299 and .410 marks across the four specifications considered there; all of these estimates are significant at the 1% level. This is reassuring as it is evidence against interaction effects between the possibly evolving difficulty level of the national exams and any ability differences between students in treatment and control groups. If, for instance, national exams get easier when the reform is introduced (as suggested from the analysis of the raw data) and if high-ability private-school students also respond better to such possibly easier exams, then this could generate misleading evidence of relatively lower achievement in public schools. The results would also suggest higher grade inflation in public schools to the extent that the internal results do not change at the same time (as, again, may be the case from the analysis of the raw data). However, the results presented above are evidence of increasing grade inflation across

the board, not only for high-ability students. Moreover, our findings of lower achievement and higher inflation also arise when focusing on public schools in the islands, where the ability interaction argument presumably does not apply.

Overall, the results indicate that the onset of individual teacher incentives led to a decrease in student achievement (when measured by national exams) and an increase of grade inflation. According to our theoretical discussion, these empirical results are consistent with incentives-related disruption in collaborative work in schools, once teachers are facing tournaments for promotions, and as internal (teacher-determined) results carry a considerable weight in final marks, thus enhancing a teacher's chances of promotion.

5 Robustness

5.1 Common trends

One important test of the strength of a causal interpretation of DID estimates concerns common trends. Indeed, if there are no interactions between treatment and other variables, as assumed for identification purposes, one would expect parallel movement between the treatment and control groups before the treatment began. We conduct this test here by considering more flexible versions of equations 1 and 2. Specifically, we allow the difference in the outcomes between treatment and control groups to vary during the period prior to the intervention. If our earlier estimates are indeed capturing a causal effect, then we expect that there will be no statistically significant differences in trends between the two groups until the occurrence of the treatment. Moreover, we also allow the effect of the education reform to vary over the $After$ period. This serves as another robustness test, as it allows us to investigate any cumulative effects of the reform.

In this context, the first equation we estimate is as follows:

$$y_{ijt} = \alpha_0 + \alpha_1 Continent_j + \sum_{k=2002}^{2011} \delta_k I(year_t = k) + \sum_{k=2002}^{2011} \gamma_k Continent_j \times I(year_t = k) + u_{ijt}. \tag{3}$$

All variables have the same meaning as before; and $I()$ is the indicator function. The parameters of interest are now the γ_k (k=2002, ..., 2005, 2007, ..., 2011), which will indicate any differences in the yearly effects for the treatment group with respect to the benchmark year (2006). As before, we consider specifications without any controls (column 1) or with

school or school-exam fixed effects (columns 2 and 3, respectively).

Table B3 (left panel) presents results based on considering internal grades as the dependent variable. We find that, across all specifications, there are no differences in trends between public schools in the islands and those in the mainland apart from one coefficient in our most restrictive specification (column 3). We also find that there are significant treatment effects in 2007, but not later. In terms of external results – Table B4 (left panel) –, we find again no evidence of different trends between the two types of schools in the baseline and school fixed effects specifications; however, two pre-reform estimates in the third column are statistically different from zero. When external grades are considered as the dependent variable, we find significant negative impacts of the incentives on public schools in the continent throughout the post-reform period.[16] These results are also displayed in Figure 2 (left panel). When a standardized external score is the regressand, we find no evidence of systematic differences in trends between both groups of students (see Table B5, left panel).

The analysis of grade inflation is again consistent with the earlier findings. Table B6 (left panel) indicates no systematic differences between the two types of schools until 2008 (except for two pre-reform years in the third specification), when grade inflation effects jump in magnitude and become statistically significant in all cases. Before that, in 2007, point estimates are already typically higher than before. In any case, grade inflation jumps even further in 2009, and especially in 2011, when all effects are significant at the 1% level (see the right panel in Figure 2).

Overall, we regard these results – particularly those of the baseline DID model and of that which controls for school fixed effects – as supportive of a causal interpretation for our main findings. Furthermore, the cumulative nature of the effects is also consistent with the cumulative nature of the reform, in the sense that the cohorts that sit their exams later (in 2008 rather than in 2007, for instance) are also typically cohorts that have been exposed to the treatment for a longer period.

We also test the common trends assumption (and the cumulativeness of the effects) in

[16] Here we also find examples of pre-reform years in which national exams exhibit significantly different means (compared to 2006) but without correspondence in terms of a differential effect between continent and islands schools. This is further evidence against a spurious relationship driven by interactions between student ability and exam difficulty.

terms of the public vs private schools comparison:

$$y_{ijt} = \alpha_0^{'} + \alpha_1^{'} Public_j + \sum_{k=2002}^{2011} \delta_k I(year_t = k) + \sum_{k=2002}^{2011} \gamma_k^{'} Public_j \times I(year_t = k) + u_{ijt}. \quad (4)$$

With the exception of the earliest years (2002 and 2003) in the period considered, we generally find little evidence of statistically significant differences during the 'before' period for the four variables considered and across the three specifications estimated for each variable – see the right panels of Tables B3, B4, and B5 for the results on internal and external marks, the latter both in levels and standardized, and of Table B6 for the results on grade inflation. The only exception to this pattern is some evidence of higher inflation in 2004 and 2005, but not earlier. However, those point estimates are generally quite smaller than their 2007-2011 counterparts. Moreover, without exception, all point estimates in 2008 or later are bigger (in absolute terms) than in 2007 (even if their differences are frequently not statistically significant), which we take as further evidence of cumulative effects of the reform. Figure 3 displays the main findings of this analysis.

The relative consistency of the effects on external marks over the 'after' period – in both control groups – is also evidence against any possible one-off disruption across public schools or amongst their teachers that coincided with exam time, even if we are unaware of any example of such an event.

5.2 Competition for promotions

According to our theoretical discussion in Section 2, the negative effects on different measures of student achievement documented above would be driven by a combination of decreased cooperation amongst teachers and increased administrative workload, both of which would shift resources away from teaching, with a potentially detrimental effect upon student learning.

This subsection offers some indirect evidence about the importance of these mechanisms. Since we do not have access to information about teachers' engagement in non-teaching school activities or other direct proxies for their collaborative work and bureaucratic workload, we focus on a testable implication of the reform. In particular, we test whether a higher *intensity* of competition for promotions under the reform's tournament component was associated with larger effects (in absolute terms) on student outcomes.

We exploit the Ministry of Education's longitudinal administrative student-teacher-school

matched dataset, MISI, to identify the tenured teachers in each continent public school who were immediately exposed to the tournament, due to their advanced rank in the old pay scale.[17] These are the teachers for whom this aspect of the reform was perhaps most salient. Hence, we would expect them to have adjusted their behaviour (e.g., collaborative approach, engagement in non-teaching activities) more strongly as a response to the new incentives scheme. Specifically, we use the teacher-school matched panel to identify the teachers that were, in 2006/07, assigned to one of the three higher ranks (out of 10) of the old pay scale; these teachers were immediately eligible to apply for the upper pay scale and *professor titular* status upon approval of the reform. We also identify those teachers who were eventually appointed to *titular* status. Then, we compute the fraction of tenured (i.e., permanent contract) teachers in each school who were eligible for (and who managed to achieve) the upper pay scale; and merge this information with the student-exam dataset used in the remainder of this paper. Finally, we estimate equations of the form:

$$y_{ijt} = \beta_0'' + \beta_1'' Intensity_j + \beta_2'' After_t + \beta_3'' Intensity_j \times After_t + u_{ijt}. \tag{5}$$

Results are shown in Table B7. We focus on external grades (left panel) and grade inflation (right panel) as dependent variables. $Intensity_j$ is a measure of the intensity of competition for promotions in each public school j – either the share of 'senior' teachers (i.e., those directly exposed to the tournament), in the top panel, or the difference between the proportions of 'senior' and 'promoted' teachers (i.e., the difference between those who were potentially eligible for *professor titular* status and those who were indeed assigned to the upper pay scale), in the bottom panel. The latter is, in practice, the fraction of senior teachers who did not manage to be assigned to the upper pay scale, in school j, in the 2006/07 school year.

We find that, after the reform, a 10 percentage points increase in the proportion of 'senior' teachers was associated with a .043 to .089 marks reduction in external exam scores and a .061 to .105 marks increase in grade inflation. All estimates are statistically significant at least at the 5% level. We obtain similar point estimates in most specifications using our second measure of intensity of competition for promotions, as reflected in the bottom panel of Table B7. However, these are not statistically significant. On the other hand, considering

[17] See Ferreira & Martins (2023) for a recent study that leverages this dataset for an analysis of upper-secondary education outcomes. Note that private schools and public schools in the island regions do not report these data to the Ministry of Education.

the standardized external score leads to negative and statistically significant point estimates using both proxies for intensity of competition.

Overall, our results suggest that the negative effects of the reform on achievement were larger (in absolute terms) in schools where the teaching staff was particularly exposed to the tournament. This is consistent with a disruption in collaborative work due to competition for promotions among colleagues, and perhaps with higher administrative workload, given that teachers in the upper pay scale were expected to contribute to a whole range of non-teaching duties (including the evaluation of their lower-ranked peers).[18]

5.3 Compositional effects

Due to the nature of our analysis, we rely on repeated cross-sectional data for our difference-in-difference estimations. Hence, if the reform had induced changes in the composition of the treatment and control groups (e.g., differential drop out rates amongst these groups and on account of certain unobserved characteristics), then our estimates could be biased. In this subsection, we leverage MISI data – in particular, its student-school matched panel – to address this possibility; test the robustness of our results to different definitions of our analysis sample; and discuss how a contemporaneous upper-secondary curriculum reform might have changed the composition of academic track cohorts in public schools.

The MISI student-school matched panel is available from 2006/07 (the school year during which the reform was approved) onward. It includes comprehensive individualized information regarding the demographic, socio-economic, and educational characteristics of the student population. Figure A2 allows for a graphical cohort-by-cohort comparison of public and private school students, enrolled in 11^{th} or 12^{th} grades, with respect to key variables with virtually no missing values: gender, age, and status as (partial or full) school welfare support beneficiary. We find that the proportion of female students in the potential sample falls year-on-year for both types of schools (with the exception of 2009, for private schools). Between 2006 and 2011, girls went from 59.4% to 57.1% of the 11^{th} and 12^{th} public-school student population (54.7% to 52.9% in private schools). Meanwhile, the average age of enrolled students (17.86 years, $SD = .841$) was virtually unchanged for both types of school

[18] Martins (2010), exploiting a 2008 law that allowed public-sector early retirements while imposing a hefty penalty per year of early retirement, finds that teachers' take-up of this option was higher in the schools that experienced a larger post-incentives-reform decline in student achievement. This is suggestive evidence of, or at least consistent with, decreased job satisfaction as a result of the new teacher incentives scheme. See Green & Heywood (2008) for a study of the correlation between performance pay and job satisfaction.

– falling .061 and .107 years, respectively, in the public and private sectors. This variation was essentially driven by a negligible decline in the proportion of 12th grade students in our population of interest. This result is important, in the sense that it suggests there seems to have been no systematic change in grade retention practices throughout the period that could affect the composition of the potential sample during and after the reform. Finally, the share of students who were full or partial beneficiaries of school welfare support increased over the period in both types of school.[19] Considering that receiving school welfare support provides a common proxy for student or household socio-economic status, the similar variation observed for both groups is reassuring for our analysis.

Furthermore, while upper-secondary school enrollments were fairly constant throughout the post-reform period (see Figure A3), it is still possible that the reform and/or an unrelated simultaneous event might have changed students' decisions in a way that would exclude them from our preferred sample (e.g., the share of students who decide not to apply for university admission even before sitting national exams may have changed throughout the period). To account for that, we conducted a series of robustness analyses using different subsets of our original student-exam data. First, we extended the range of data examined from first-call results (which account for over 70% of the total number of exams) to first- and second-calls. Then, we extended the range of data considered even further, thereby also including resit students and those not applying for university entry. Ultimately, our estimated effects on external scores were robust to the consideration of all available student-exam observations (see Table B2).

Finally, the introduction and rapid expansion of upper-secondary vocational (or VET) courses in public schools from 2004 onward, both in the mainland and the islands, led to substantial change – contemporaneous with the incentives reform – for Portugal's education system. Just between 2006 and 2010, the proportion of upper-secondary education students who were enrolled in VET courses increased from 13.1% to 31.4%. Although many of these early VET students would have been likely to drop out in the absence of a nonacademic track,[20] the introduction of VET courses in public schools also led to 'displacement' from the

[19]There is a large increase in school welfare support take-up after 2009 as a result of a reform to its provision. Private school students may receive welfare support provided their schools receive public funding to admit a given number of classes free of charge (*Contratos de Associação*; similar to charter schools). Full welfare support entitles students to, for instance, free school meals and textbooks.

[20]Before a 2009 reform, with effect in 2012, pupils were only required to stay in school up to the age of 15 (i.e., until the end of 9th grade or lower-secondary school).

academic track (Ferreira & Martins 2023). Notice that VET students are not required to sit national exams in order to graduate, so they are absent from our preferred analysis sample. Therefore, important unobserved features of the exam-taking population (in public schools) might have changed throughout the period considered in our analysis, which would cast doubt upon our DID estimates using the private schools comparison group.

However, all evidence suggests that the VET track appealed mostly to students from less-privileged socio-economic and educational backgrounds, and with worse prior academic achievement (measured either by their results in earlier national exams or previous experiences of grade retention). Therefore, the apparent displacement of academically-weaker public-school students from the academic track should have *increased* its students' results relative to private schools, particularly in the later years of the period under study. Nevertheless, we find the opposite: our estimated effects of the reform appear to be cumulative and stronger for later cohorts. Hence, on these grounds, our results are more likely *conservative*, attenuated estimates of the reform's effects than the opposite.

5.4 Other control variables and tests

Student achievement is affected by many variables other than those related to teacher merit pay. In particular, socio-economic variables may matter greatly. Given the non-experimental setting of our analysis, it is not impossible (even if unlikely, given the evidence produced so far) that the different types of schools that we contrast experience different trends in such socio-economic variables which just happen to coincide with the introduction of the new teacher incentives.

In order to assess the empirical content of this alternative view, we add to our specifications (equations 1 and 2) different characteristics of the local labour market of each school that will proxy for the socio-economic environment of these students. Specifically, we draw on the *Quadros de Pessoal* (QP) matched employer-employee data set, which reports detailed firm-, establishment-, and worker-level information of all firms in Portugal that employ at least one worker (see Ferreira & Martins (2023) for more detail about these data). We focus on the establishment- and worker-level dimensions, as this allows us to compute region-year characteristics at the most detailed level of aggregation available on the JNE data (the *concelho* level). The QP variables we add to our student-level equations are the (log) mean monthly

wage, the female ratio, the average schooling attainment, and the (log) total number of workers. These variables are computed from all workers employed in the same *concelho* where the student's school is located and in the same year to which the student's results refer.[21] The results again present strong evidence of lower achievement in terms of national exams and increasing grade inflation, for both control groups (see Table B8).

We have also conducted a number of additional robustness tests. Specifically, we studied possible differences from the benchmark results alternatively in urban areas, in large schools, or in core subjects only. In addition, we controlled for some student characteristics, namely age, gender, and grade (11^{th} or 12^{th}), available in the JNE data, although only from 2006 onward. In all cases, the qualitative results across the different specifications were unchanged and only relatively minor differences were found in terms of the quantitative findings.[22]

6 Conclusions

There is great public policy interest in understanding the role of worker incentives in improving public service delivery. A particular area in which this issue is very important is that of teacher incentives, with a view to improving student achievement. This paper sheds light on this question by examining the introduction of performance-related pay in all public schools in Portugal. Our approach is based on a difference-in-differences analysis drawing on upper-secondary national exams data and two complementary control groups. These control groups either were exposed to a lighter version of the intervention (public schools in two autonomous regions) or were not exposed at all (private schools).

Our results consistently indicate that the increased focus on individual teacher performance caused a significant decline in student achievement, as measured by national exams. However, the decline in achievement is smaller or virtually zero when considering marks set by teachers. The two results together suggest that grade inflation was another consequence

[21] We match QP data of year t to JNE data of year $t+1$, given that the JNE data concern academic years that begin in September of year t to June of year $t+1$ and the QP data refers to October of each year.

[22] To address concerns about external validity that have been directed towards case-study or experimental settings, Martins (2010) sought to understand the dispersion of effect estimates across different treatment schools. In particular, equations 1 and 2, augmented with school fixed effects and a control for the number of exams in each school-year, were re-estimated. Each continental public school was separately compared against either all public/island schools or all private schools. This approach generated as many DID estimates as the number of treatment-group schools, from which measures of their dispersion were computed. The results suggested a considerable scope for variation of the effects across different schools; for instance, taking external exam scores as the outcome of interest, the estimated impact of the reform on about one fourth of continental public schools would have been positive. However, this analysis at a higher level of aggregation yielded results consistent with the main findings.

of this reform. This view is supported by our triple-difference evidence and is consistent with the emphasis placed on student results by the new promotion criteria. Furthermore, we find additional support for a causal interpretation of our results from our analysis of common trends; also, estimates prove to be robust to different control variables and different data subsets. The analysis of 'competition for promotions' across public schools also supports the theoretical mechanisms (and much anecdotal evidence) that predict the empirical results, namely disruption of teacher cooperation created by tournaments for promotions and increased administrative workloads, both potentially resulting in worse student outcomes.

On a methodological note, our use of official, administrative population data should ensure greater reliability in terms of the external validity of the findings and other potential problems such as measurement error bias and Hawthorne effects. As we examine a period of five years after the reform was introduced, which is longer than most other related studies, and find consistent negative achievement effects, our analysis is not picking up implementation problems that may otherwise erode over time.

While our results are negative regarding the value of the specific reform examined here, the findings are useful for public policy. The results highlight the potential negative effects of individual incentives in the public sector or at least in public schools.

References

Aaronson, D., Barrow, L. & Sander, W. (2007), 'Teachers and student achievement in the Chicago public high schools', *Journal of Labor Economics* **25**, 95–135.

Atkinson, A., Burgess, S., Croxson, B., Gregg, P., Propper, C., Slater, H. & Wilson, D. (2009), 'Evaluating the impact of performance-related pay for teachers in England', *Labour Economics* **16**(3), 251–261.

Bandiera, O., Barankay, I. & Rasul, I. (2005), 'Social preferences and the response to incentives: Evidence from personnel data', *Quarterly Journal of Economics* **120**(3), 917–962.

Barlevy, G. & Neal, D. (2012), 'Pay for percentile', *American Economic Review* **102**(5), 1805–1831.

Barrera-Osorio, F. & Raju, D. (2017), 'Teacher performance pay: Experimental evidence from Pakistan', *Journal of Public Economics* **148**, 75–91.

Behrman, J. R., Parker, S. W., Todd, P. E. & Wolpin, K. I. (2015), 'Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools', *Journal of Political Economy* **123**(2), 325–364.

Bergman, P. & Hill, M. J. (2018), 'The effects of making performance information public: Regression discontinuity evidence from Los Angeles teachers', *Economics of Education Review* **66**, 104–113.

Brehm, M., Imberman, S. A. & Lovenheim, M. F. (2017), 'Achievement effects of individual performance incentives in a teacher merit pay tournament', *Labour Economics* **44**, 133–150.

Burgess, S., Propper, C., Ratto, M. & Tominey, E. (2017), 'Incentives in the public sector: Evidence from a government agency', *Economic Journal* **127**(605), 117–141.

Chetty, R., Friedman, J. N. & Rockoff, J. E. (2014), 'Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood', *American Economic Review* **104**(9), 2633–79.

Dee, T. S. & Wyckoff, J. (2015), 'Incentives, selection, and teacher performance: Evidence from IMPACT', *Journal of Policy Analysis and Management* **34**(2), 267–297.

Dixit, A. (2002), 'Incentives and organizations in the public sector: An interpretative review', *Journal of Human Resources* **37**(4), 696–727.

Eren, O. (2019), 'Teacher incentives and student achievement: Evidence from an advancement program', *Journal of Policy Analysis and Management* **38**(4), 867–890.

European Commission (2007), The education system in Portugal, Eurybase report.

Fehr, E. & Schmidt, K. M. (1999), 'A theory of fairness, competition, and cooperation', *Quarterly Journal of Economics* **114**(3), 817–868.

Fehr, E. & Schmidt, K. M. (2004), 'Fairness and incentives in a multi-task principal-agent model', *Scandinavian Journal of Economics* **106**(3), 453–474.

Ferreira, J. R. & Martins, P. S. (2023), Can Vocational Education Improve Schooling and Labour Outcomes? Evidence from a Large Expansion, IZA Discussion Paper 16474.

Figlio, D. N. & Kenny, L. W. (2007), 'Individual teacher incentives and student performance', *Journal of Public Economics* **91**(5-6), 901–914.

Fryer, R. G. (2013), 'Teacher incentives and student achievement: Evidence from New York City public schools', *Journal of Labor Economics* **31**(2), 373–407.

Fryer, R. G., Levitt, S. D., List, J. & Sadoff, S. (2022), 'Enhancing the efficacy of teacher incentives through framing: A field experiment', *American Economic Journal: Economic Policy* **14**(4), 269–99.

Glewwe, P., Ilias, N. & Kremer, M. (2010), 'Teacher incentives', *American Economic Journal: Applied Economics* **2**(3), 205–27.

Goodman, S. F. & Turner, L. J. (2013), 'The design of teacher incentive pay and educational outcomes: Evidence from the New York City bonus program', *Journal of Labor Economics* **31**(2), 409–420.

Green, C. & Heywood, J. S. (2008), 'Does performance pay increase job satisfaction?', *Economica* **75**(300), 710–728.

Imberman, S. A. & Lovenheim, M. F. (2015), 'Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system', *Review of Economics and Statistics* **97**(2), 364–386.

Jacob, B. A. & Lefgren, L. (2008), 'Can principals identify effective teachers? Evidence on subjective performance evaluation in education', *Journal of Labor Economics* **26**, 101–136.

Jacob, B. A. & Levitt, S. D. (2003), 'Rotten apples: An investigation of the prevalence and predictors of teacher cheating', *Quarterly Journal of Economics* **118**(3), 843–877.

Jones, M. D. (2013), 'Teacher behavior under performance pay incentives', *Economics of Education Review* **37**, 148–164.

Kane, T. J. & Staiger, D. O. (2002), 'The promise and pitfalls of using imprecise school accountability measures', *Journal of Economic Perspectives* **16**(4), 91–114.

Lavy, V. (2002), 'Evaluating the effect of teachers' group performance incentives on pupil achievement', *Journal of Political Economy* **110**(6), 1286–1317.

Lavy, V. (2009), 'Performance pay and teachers' effort, productivity, and grading ethics', *American Economic Review* **99**(5), 1979–2011.

Lazear, E. P. (2000), 'Performance pay and productivity', *American Economic Review* **90**(5), 1346–1361.

Lazear, E. P. (2003), 'Teacher incentives', *Swedish Economic Policy Review* **10**(2), 179–214.

Lazear, E. P. & Rosen, S. (1981), 'Rank-order tournaments as optimum labor contracts', *Journal of Political Economy* **89**(5), 841–64.

Leone, T. (2024), 'Does a productivity bonus pay off? The effects of teacher-incentive pay on student achievement in Brazilian schools', *Economic Development and Cultural Change* **72**(3), 1317–1356.

Loyalka, P., Sylvia, S., Liu, C., Chu, J. & Shi, Y. (2019), 'Pay by design: Teacher performance pay design and the distribution of student achievement', *Journal of Labor Economics* **37**(3), 621–662.

Martins, P. S. (2008), 'Dispersion in wage premiums and firm performance', *Economics Letters* **101**(1), 63–65.

Martins, P. S. (2010), Individual Teacher Incentives, Student Achievement and Grade Inflation, Centre for the Economics of Education Discussion Paper 112.

Martins, P. S. (2023), The Wage Effects of Employers' Associations: A Case Study of the Private Schools Sector, IZA Discussion Papers 16476.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C. & Rajani, R. (2019), 'Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania', *Quarterly Journal of Economics* **134**(3), 1627–1673.

Meyer, B. D. (1995), 'Natural and quasi-experiments in economics', *Journal of Business & Economic Statistics* **13**(2), 151–61.

Muralidharan, K. & Sundararaman, V. (2011), 'Teacher performance pay: Experimental evidence from India', *Journal of Political Economy* **119**(1), 39–77.

Neal, D. (1997), 'The effects of Catholic secondary schooling on educational achievement', *Journal of Labor Economics* **15**(1), 98–123.

Nunes, L. C., Reis, A. B. & Seabra, C. (2015), 'The publication of school rankings: A step toward increased accountability?', *Economics of Education Review* **49**, 15–23.

OECD (2001), 'Education at a glance 2001', OECD Directorate for Education, Paris.

Pereira dos Santos, J., Tavares, J. & Mesquita, J. (2021), 'Leave them kids alone! National exams as a political tool', *Public Choice* **189**, 405–426.

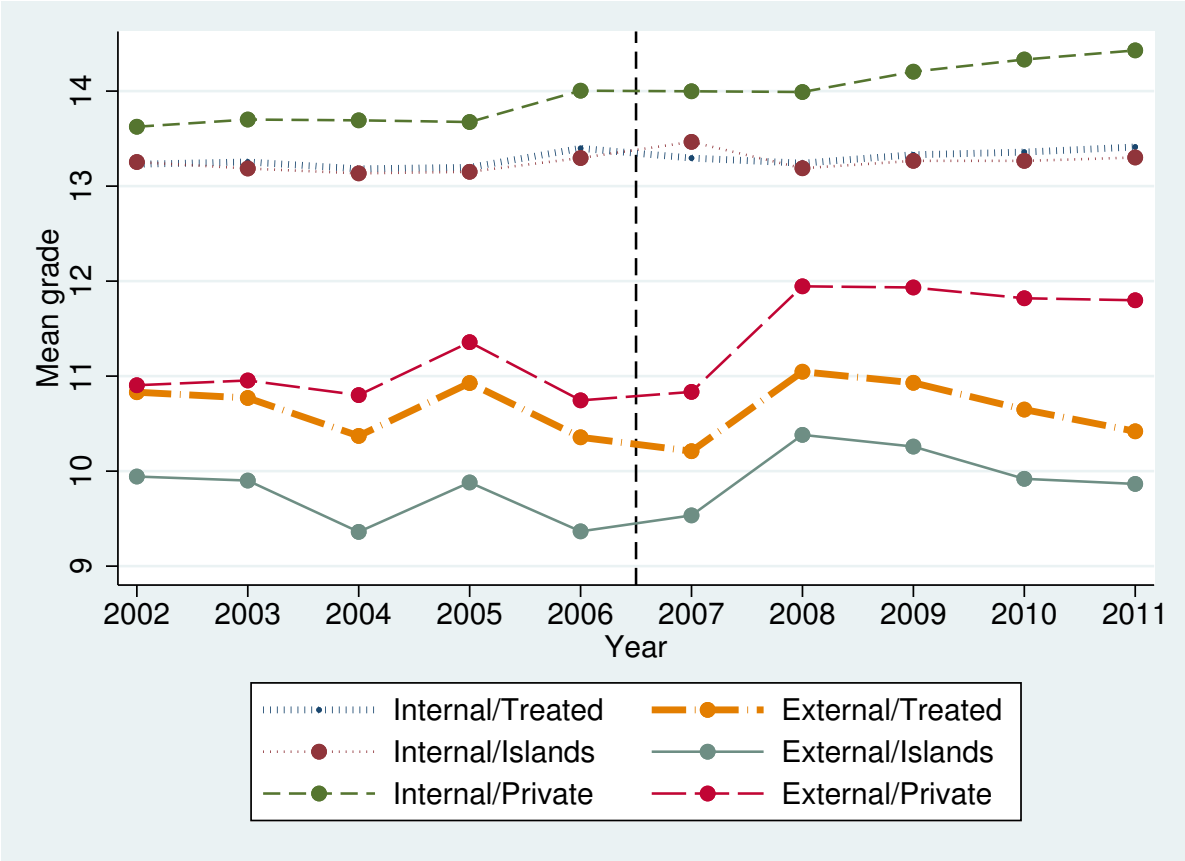
Pham, L. D., Nguyen, T. D. & Springer, M. G. (2021), 'Teacher merit pay: A meta-analysis', *American Educational Research Journal* **58**(3), 527–566.

Rivkin, S. G., Hanushek, E. A. & Kain, J. F. (2005), 'Teachers, schools, and academic achievement', *Econometrica* **73**(2), 417–458.

Sojourner, A. J., Mykerezi, E. & West, K. L. (2014), 'Teacher pay reform and productivity: Panel data evidence from adoptions of Q-Comp in Minnesota', *Journal of Human Resources* **49**(4), 945–981.

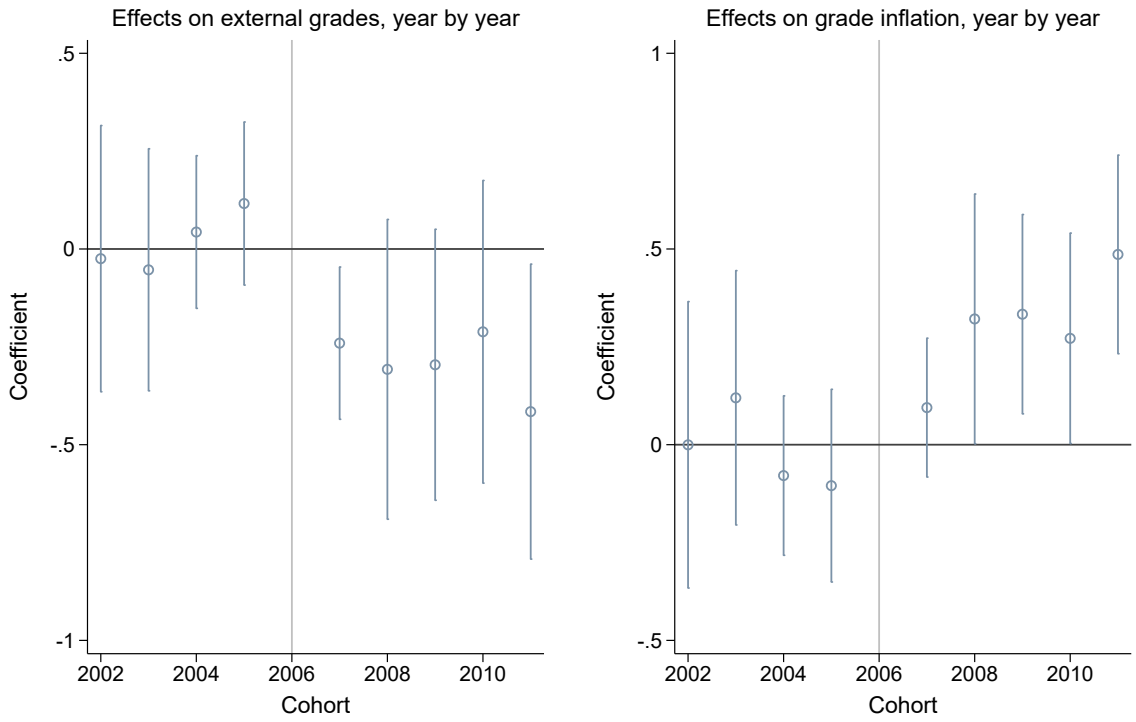
Figures and Tables

Figure 1: **Internal and external grades across groups and time**



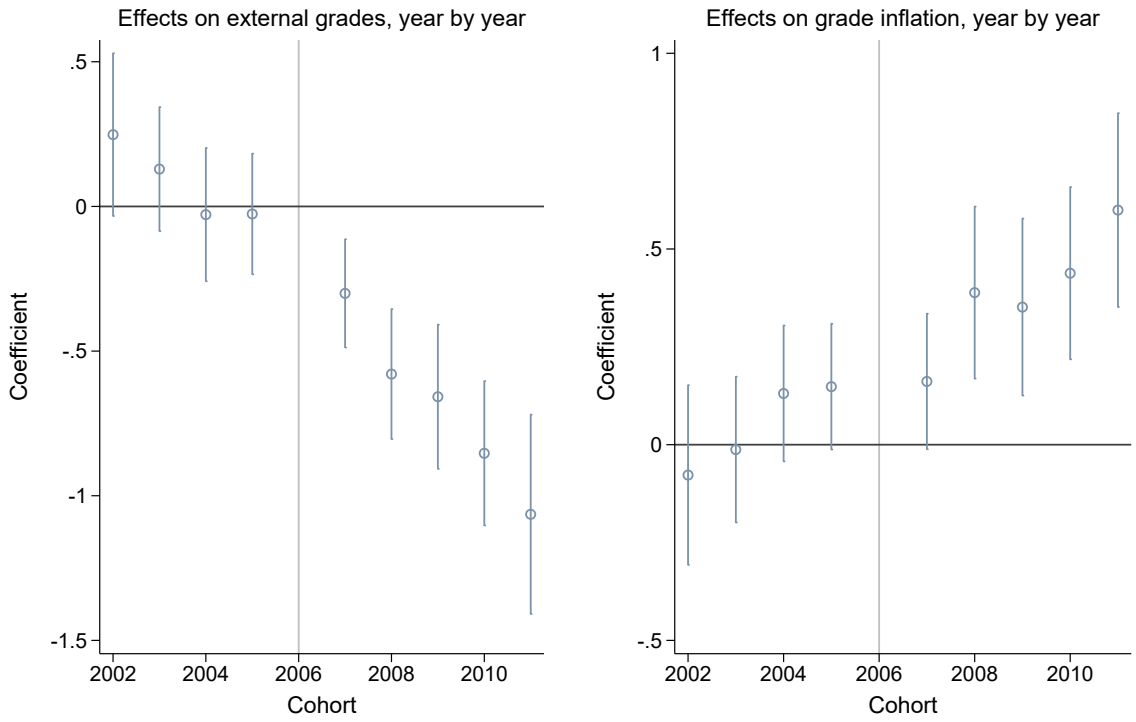
Source: Authors' calculations based on JNE data. Mean internal and external marks of students by year and type of school (public schools in continental Portugal – 'Treated'; public schools in the Azores and Madeira – 'Islands'; and private schools in continental Portugal – 'Private'). The vertical dashed line indicates the introduction of the teacher incentives reform.

Figure 2: **Difference-in-difference estimates (Islands control group)**



Notes: Author's estimates based on JNE data and reported in more detail in Column 1 of Tables B4 and B6. 95% confidence intervals are reported.

Figure 3: **Difference-in-difference estimates (Private control group)**



Notes: Author's estimates based on JNE data and reported in more detail in Column 4 of Tables B4 and B6. 95% confidence intervals are reported.

Table 1: **Descriptive Statistics**

Variable	Mean	Std. Dev.	Min.	Max.	Obs.
School-level data					
Internal score	13.291	0.729	10.5	18.167	6101
External score	10.418	1.332	2.7	16.224	6101
Internal - External score	2.694	1.065	-3.286	9	6101
Public	0.806	0.395	0	1	6101
Continent	0.943	0.233	0	1	6058
No. Exams	331.979	264.196	1	2386	6101
School w/ exams every year	0.913	0.281	0	1	6101
Exam-level data					
Internal score	13.366	2.594	10	20	2025402
External score	10.682	3.973	0	20	2025401
Internal - External score	2.502	3.1	-10	19	2025401
Public	0.881	0.324	0	1	2025402
Continent	0.948	0.222	0	1	2019083
No. Exams	542.198	327.644	1	2386	2025402
School w/ exams every year	0.968	0.177	0	1	2025402
2002	0.134	0.34	0	1	2025402
2003	0.124	0.33	0	1	2025402
2004	0.087	0.281	0	1	2025402
2005	0.097	0.296	0	1	2025402
2006	0.117	0.321	0	1	2025402
2007	0.08	0.272	0	1	2025402
2008	0.083	0.276	0	1	2025402
2009	0.094	0.292	0	1	2025402
2010	0.092	0.289	0	1	2025402
2011	0.092	0.289	0	1	2025402
Maths (12)	0.147	0.354	0	1	2025402
Portuguese (12)	0.148	0.355	0	1	2025402
History (12)	0.047	0.211	0	1	2025402
Biology & Geology (11)	0.137	0.343	0	1	2025402
Physics & Chemistry (11)	0.133	0.34	0	1	2025402

Notes: Authors' calculations based on *Júri Nacional de Exames* data. The internal (external) score refers to the mark obtained by each student in each module at the school (national exam) level. 'Public' and 'Continent' are dummy variables which are equal to one for students in public schools or schools located in mainland Portugal, respectively. There are 682 schools, of which 562 are observed in all ten years ('School w/ exams every year'), resulting in 6,101 school-year observations and 2,025,402 exam-level observations.

Table 2: **Effects on internal grades**

A. Islands control group	(1)	(2)	(3)	(4)
After	0.118	0.181	0.186	0.178
	(0.110)	(0.098)*	(0.107)*	(0.094)*
Continent	0.045	0.072		
	(0.078)	(0.091)		
Continent-After	-0.009	-0.032	-0.083	-0.112
	(0.111)	(0.100)	(0.107)	(0.096)
Obs.	1,573,914	1,573,914	1,535,967	920,577
Mean dep. var.	13.275	13.275	13.278	13.170
Mean dep. var. (Islands)	13.237	13.237	13.258	13.149
R^2	0.000	0.004	0.024	0.048

B. Private control group	(1)	(2)	(3)	(4)
After	0.552	0.634	0.501	0.552
	(0.091)***	(0.103)***	(0.095)***	(0.118)***
Public	-0.430	-0.521		
	(0.141)***	(0.149)***		
Public-After	-0.443	-0.482	-0.411	-0.487
	(0.093)***	(0.103)***	(0.100)***	(0.119)***
Obs.	1,695,840	1,695,840	1,641,336	989,489
Mean dep. var.	13.358	13.358	13.357	13.260
Mean dep. var. (Private)	13.933	13.933	13.957	13.920
R^2	0.009	0.012	0.045	0.073
Log No. Exams	No	Yes	Yes	No
School FE	No	No	Yes	No
School-Exam FE	No	No	No	Yes

Notes: Dependent variable is the school-level grade of each student in each exam in each year. Dummy $After$ is one for 2007-2011 only. Data used: 2002 to 2011, except for 2006. Columns 2-3 include a control for the (log of the) number of exams taken in each school in each year. Column 3 controls for school fixed effects; and column 4 controls for school-subject fixed effects (the 5 main subjects are considered in this column: Portuguese, Maths, History, Biology & Geology, and Physics & Chemistry). Robust standard errors, allowing for clustering at the school level. Significance levels: *: 0.1; **: 0.05; ***: 0.01.

Table 3: **Effects on external grades**

A. Islands control group	(1)	(2)	(3)	(4)
After	0.188	0.346	0.258	0.934
	(0.131)	(0.115)***	(0.128)**	(0.148)***
Continent	0.957	1.024		
	(0.148)***	(0.147)***		
Continent-After	-0.296	-0.355	-0.447	-0.702
	(0.135)**	(0.120)***	(0.130)***	(0.152)***
Obs.	1,573,913	1,573,913	1,535,966	920,576
Mean dep. var.	10.661	10.661	10.674	10.233
Mean dep. var. (Islands)	9.897	9.897	9.946	9.358
R^2	0.002	0.011	0.046	0.090

B. Private control group	(1)	(2)	(3)	(4)
After	0.696	0.876	0.501	1.098
	(0.122)***	(0.148)***	(0.132)***	(0.182)***
Public	-0.232	-0.431		
	(0.158)	(0.183)**		
Public-After	-0.804	-0.889	-0.705	-0.866
	(0.126)***	(0.151)***	(0.137)***	(0.185)***
Obs.	1,695,839	1,695,839	1,641,335	989,488
Mean dep. var.	10.785	10.785	10.794	10.378
Mean dep. var. (Private)	11.345	11.345	11.402	11.081
R^2	0.004	0.012	0.057	0.103
Log No. Exams	No	Yes	Yes	No
School FE	No	No	Yes	No
School-Exam FE	No	No	No	Yes

Notes: Dependent variable is the the national exam grade of each student in each exam in each year. Dummy *After* is one for 2007-2011 only. Data used: 2002 to 2011, except for 2006. Columns 2-3 include a control for the (log of the) number of exams taken in each school in each year. Column 3 controls for school fixed effects; and column 4 controls for school-subject fixed effects (the 5 main subjects are considered in this column: Portuguese, Maths, History, Biology & Geology, and Physics & Chemistry). Robust standard errors, allowing for clustering at the school level. Significance levels: *: 0.1; **: 0.05; ***: 0.01.

Table 4: **Effects on grade inflation**

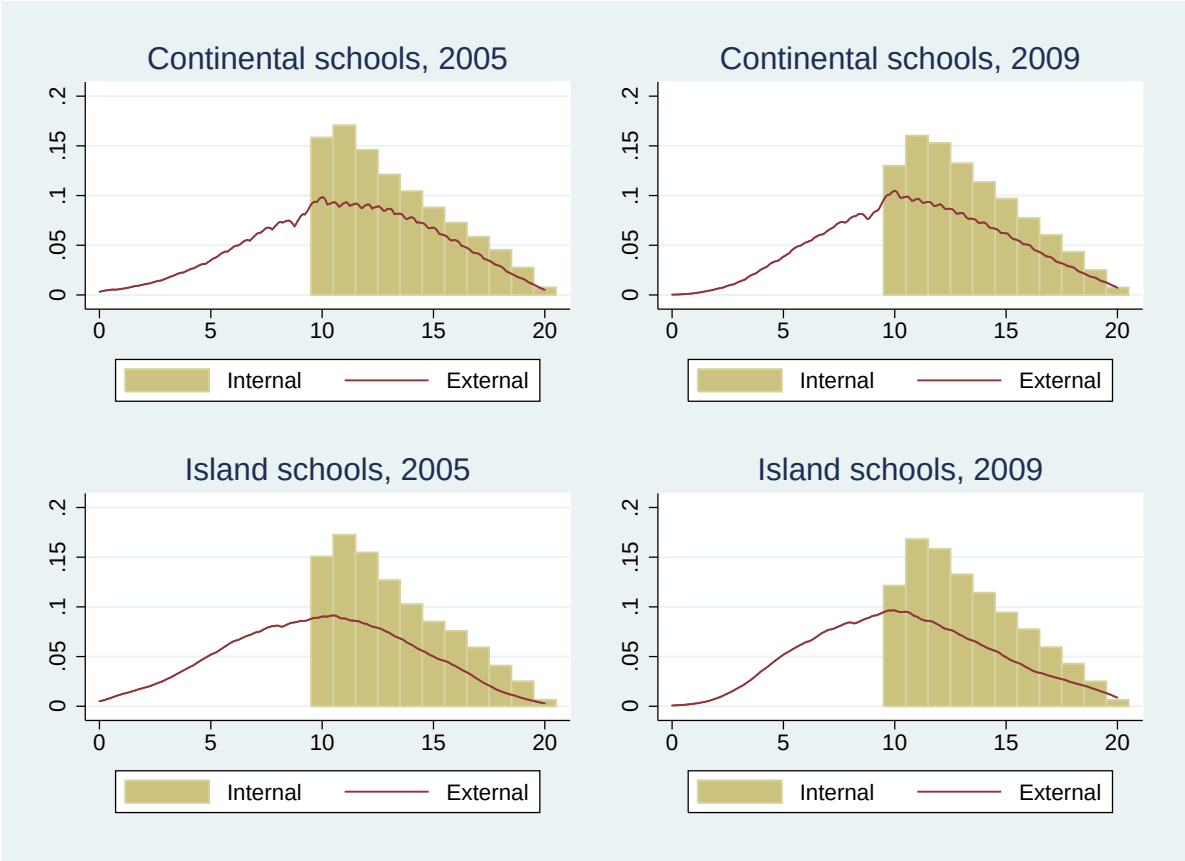
A. Islands control group	(1)	(2)	(3)	(4)
After	-0.062	-0.158	-0.065	-0.753
	(0.102)	(0.109)	(0.114)	(0.135)***
Continent	-0.941	-0.982		
	(0.124)***	(0.109)***		
Continent-After	0.299	0.335	0.376	0.599
	(0.106)***	(0.112)***	(0.118)***	(0.139)***
Obs.	1,573,913	1,573,913	1,535,966	920,576
Mean dep. var.	2.430	2.430	2.420	2.757
Mean dep. var. (Islands)	3.178	3.178	3.149	3.628
R^2	0.005	0.010	0.046	0.136

B. Private control group	(1)	(2)	(3)	(4)
After	-0.126	-0.225	0.015	-0.536
	(0.090)	(0.093)**	(0.083)	(0.114)***
Public	-0.196	-0.086		
	(0.121)	(0.148)		
Public-After	0.364	0.410	0.299	0.381
	(0.095)***	(0.099)***	(0.087)***	(0.119)***
Obs.	1,695,839	1,695,839	1,641,335	989,488
Mean dep. var.	2.388	2.388	2.377	2.701
Mean dep. var. (Private)	2.399	2.399	2.366	2.654
R^2	0.001	0.005	0.048	0.137
Log No. Exams	No	Yes	Yes	No
School FE	No	No	Yes	No
School-Exam FE	No	No	No	Yes

Notes: Dependent variable is the difference between the internal (school) grade and the external (national exam) grade of each student in each exam in each year. Dummy $After$ is one for 2007-2011 only. Data used: 2002 to 2011, except for 2006. Columns 2-3 include a control for the (log of the) number of exams taken in each school in each year. Column 3 controls for school fixed effects; and column 4 controls for school-subject fixed effects (the 5 main subjects are considered in this column: Portuguese, Maths, History, Biology & Geology, and Physics & Chemistry). Robust standard errors, allowing for clustering at the school level. Significance levels: *: 0.1; **: 0.05; ***: 0.01.

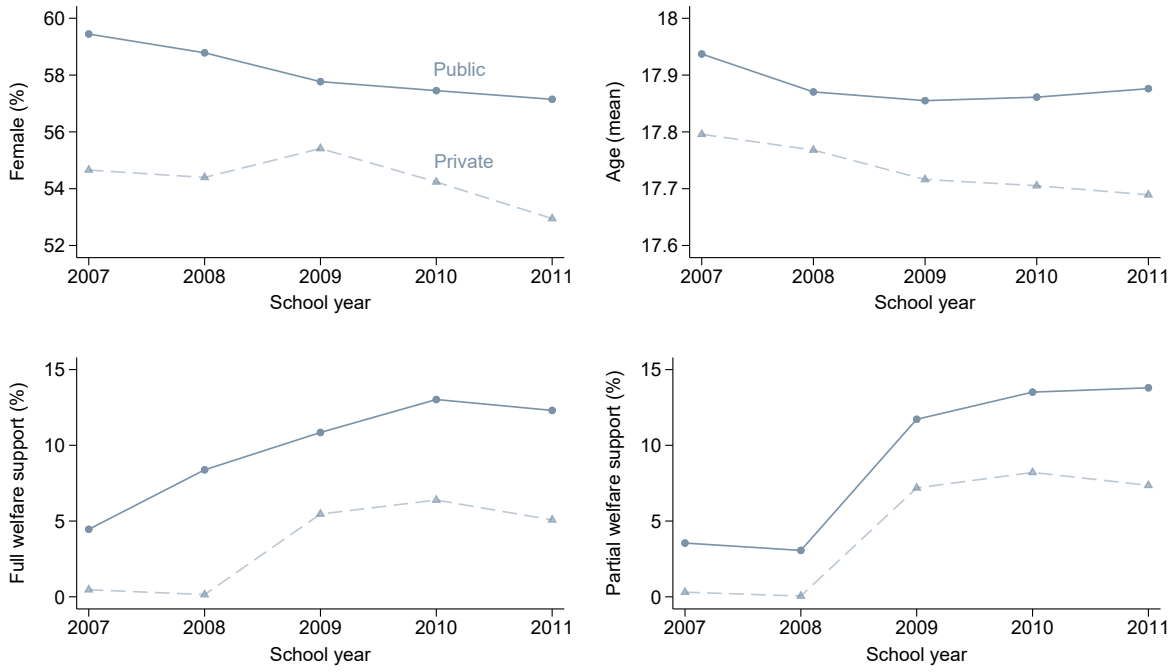
Appendix: Supplementary Figures and Tables

Figure A1: **Distribution of grades: Islands and Continent, 2005 and 2009**



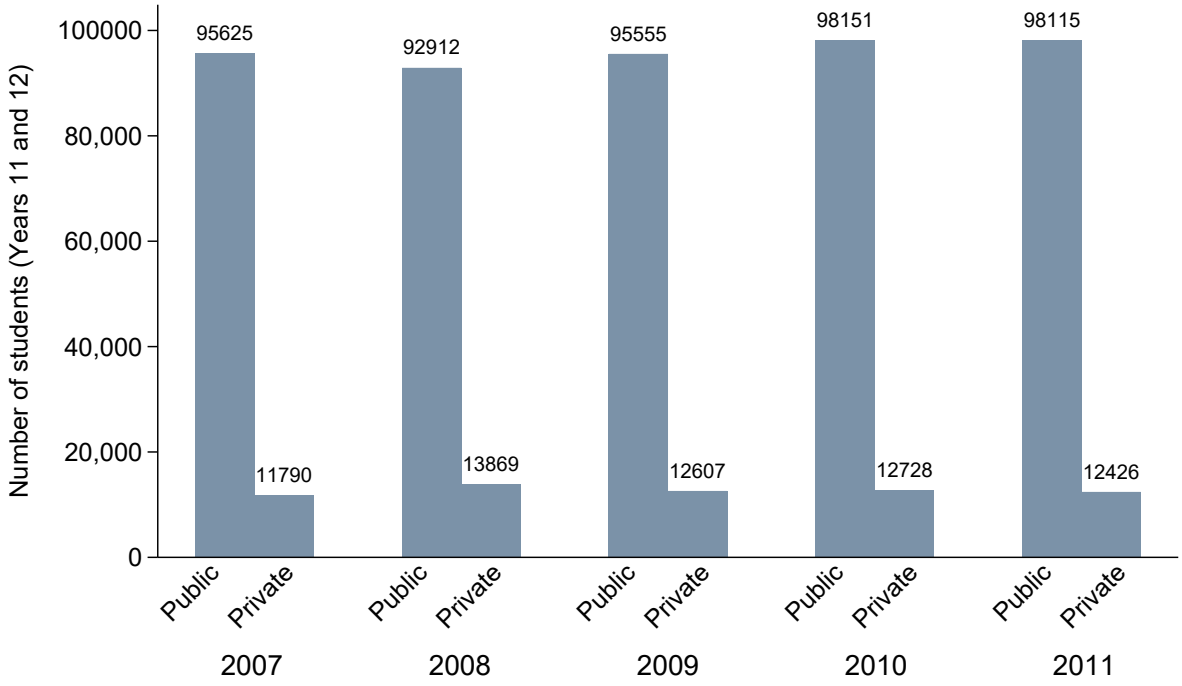
Source: Martins (2010). The external marks result from kernel estimation.

Figure A2: **Selected characteristics of (11^{th} and 12^{th} grade) public and private school students in mainland Portugal (2006/07 – 2010/11)**



Notes: The plots present (counterclockwise) the shares of female students, full welfare support beneficiaries, and partial welfare support recipients, as well as the average age, of all 11^{th} and 12^{th} grade academic-track students in mainland Portugal, by school type. Authors' computations based on MISI data.

Figure A3: **Number of 11th and 12th grade academic-track students in mainland Portugal, by school type (2006/07 – 2010/11)**



Notes: Authors' computations based on MISI data.

Table B1: **Effects on (standardized) external grades**

A. Islands control group	(1)	(2)	(3)	(4)
After	0.029	0.065	0.033	0.117
	(0.038)	(0.033)*	(0.038)	(0.038)***
Continent	0.227	0.243		
	(0.036)***	(0.037)***		
Continent-After	-0.030	-0.044	-0.068	-0.147
	(0.039)	(0.035)	(0.038)*	(0.039)***
Obs.	1,573,913	1,573,913	1,535,966	920,576
Mean dep. var. (Islands)	-0.200	-0.200	-0.186	-0.224
R^2	0.002	0.010	0.046	0.074

B. Private control group	(1)	(2)	(3)	(4)
After	0.171	0.212	0.111	0.155
	(0.030)***	(0.037)***	(0.032)***	(0.042)***
Public	-0.072	-0.118		
	(0.040)*	(0.045)***		
Public-After	-0.195	-0.215	-0.172	-0.205
	(0.031)***	(0.037)***	(0.033)***	(0.043)***
Obs.	1,695,839	1,695,839	1,641,335	989,488
Mean dep. var. (Private)	0.150	0.150	0.165	0.187
R^2	0.004	0.011	0.058	0.089
Log No. Exams	No	Yes	Yes	No
School FE	No	No	Yes	No
School-Exam FE	No	No	No	Yes

Notes: Dependent variable is the standardized national exam grade of each student in each exam in each year. Dummy $After$ is one for 2007-2011 only. Data used: 2002 to 2011, except for 2006. Columns 2-3 include a control for the (log of the) number of exams taken in each school in each year. Column 3 controls for school fixed effects; and column 4 controls for school-subject fixed effects (the 5 main subjects are considered in this column: Portuguese, Maths, History, Biology & Geology, and Physics & Chemistry). Robust standard errors, allowing for clustering at the school level. Significance levels: *: 0.1; **: 0.05; ***: 0.01.

Table B2: **Effects on external grades (robustness)**

A. Islands control group	(1)	(2)	(3)	(4)
After	0.344	0.377	0.335	1.198
	(0.111)***	(0.096)***	(0.108)***	(0.132)***
Continent	0.571	0.574		
	(0.126)***	(0.111)***		
Continent-After	-0.118	-0.129	-0.226	-0.374
	(0.114)	(0.100)	(0.110)**	(0.135)***
Obs.	3,252,138	3,252,138	3,174,960	1,991,027
Mean dep. var.	9.692	9.692	9.704	9.180
Mean dep. var. (Islands)	9.213	9.213	9.257	8.654
R^2	0.001	0.007	0.033	0.085

B. Private control group	(1)	(2)	(3)	(4)
After	0.942	1.009	0.757	1.569
	(0.112)***	(0.129)***	(0.131)***	(0.168)***
Public	-0.310	-0.498		
	(0.162)*	(0.180)***		
Public-After	-0.715	-0.764	-0.649	-0.745
	(0.115)***	(0.131)***	(0.134)***	(0.170)***
Obs.	3,498,290	3,498,290	3,381,647	2,136,528
Mean dep. var.	9.802	9.802	9.810	9.309
Mean dep. var. (Private)	10.427	10.427	10.489	10.095
R^2	0.005	0.009	0.047	0.102
Log No. Exams	No	Yes	Yes	No
School FE	No	No	Yes	No
School-Exam FE	No	No	No	Yes

Notes: Dependent variable is the national exam grade of each student in each exam in each year. The whole potential sample is considered (i.e., including second calls, external students, etc.). Dummy $After$ is one for 2007-2011 only. Data used: 2002 to 2011, except for 2006. Columns 2-3 include a control for the (log of the) number of exams taken in each school in each year. Column 3 controls for school fixed effects; and column 4 controls for school-subject fixed effects (the 5 main subjects are considered in this column: Portuguese, Maths, History, Biology & Geology, and Physics & Chemistry). Robust standard errors, allowing for clustering at the school level. Significance levels: *: 0.1; **: 0.05; ***: 0.01.

Table B3: **Effects on internal grades, year by year**

Treated group:	Continent public schools			Continent public schools		
Control group:	Azores and Madeira schools			Private schools		
	(1)	(2)	(3)	(4)	(5)	(6)
2002	-0.167	-0.166	-0.123	-0.352	-0.304	-0.286
	(0.070)**	(0.071)**	(0.074)*	(0.072)***	(0.075)***	(0.096)***
2003	-0.238	-0.237	-0.210	-0.278	-0.248	-0.227
	(0.078)***	(0.069)***	(0.083)**	(0.078)***	(0.069)***	(0.087)***
2004	-0.146	-0.110	-0.025	-0.276	-0.245	-0.242
	(0.060)**	(0.062)*	(0.100)	(0.077)***	(0.074)***	(0.090)***
2005	-0.199	-0.156	-0.104	-0.302	-0.241	-0.275
	(0.083)**	(0.083)*	(0.087)	(0.077)***	(0.070)***	(0.086)***
2007	0.013	0.049	0.089	0.005	0.008	0.080
	(0.051)	(0.055)	(0.051)*	(0.053)	(0.057)	(0.068)
2008	-0.156	-0.085	-0.032	0.053	0.067	0.115
	(0.086)*	(0.091)	(0.081)	(0.064)	(0.060)	(0.072)
2009	-0.094	-0.009	0.012	0.248	0.236	0.282
	(0.118)	(0.122)	(0.096)	(0.070)***	(0.064)***	(0.071)***
2010	-0.085	0.017	0.120	0.391	0.355	0.438
	(0.131)	(0.136)	(0.122)	(0.084)***	(0.081)***	(0.088)***
2011	-0.042	0.058	0.106	0.488	0.433	0.511
	(0.146)	(0.150)	(0.148)	(0.108)***	(0.095)***	(0.099)***
Treated	0.029			-0.560		
	(0.102)			(0.129)***		
Treated-2002	-0.015	0.003	0.055	0.170	0.152	0.219
	(0.073)	(0.074)	(0.077)	(0.075)**	(0.080)*	(0.099)**
Treated-2003	0.076	0.088	0.149	0.116	0.103	0.166
	(0.080)	(0.072)	(0.086)*	(0.080)	(0.072)	(0.090)*
Treated-2004	-0.026	-0.020	-0.048	0.104	0.092	0.170
	(0.063)	(0.065)	(0.102)	(0.079)	(0.070)	(0.092)*
Treated-2005	0.019	-0.001	0.010	0.122	0.069	0.181
	(0.085)	(0.085)	(0.090)	(0.079)	(0.069)	(0.089)**
Treated-2007	-0.150	-0.161	-0.158	-0.142	-0.151	-0.149
	(0.054)***	(0.056)***	(0.054)***	(0.055)**	(0.062)**	(0.070)**
Treated-2008	0.012	-0.046	-0.064	-0.196	-0.225	-0.211
	(0.088)	(0.091)	(0.084)	(0.066)***	(0.067)***	(0.075)***
Treated-2009	0.034	-0.048	-0.028	-0.308	-0.310	-0.298
	(0.120)	(0.122)	(0.099)	(0.073)***	(0.069)***	(0.074)***
Treated-2010	0.059	-0.033	-0.063	-0.418	-0.390	-0.382
	(0.133)	(0.136)	(0.124)	(0.086)***	(0.085)***	(0.091)***
Treated-2011	0.067	-0.023	-0.010	-0.462	-0.418	-0.415
	(0.148)	(0.150)	(0.149)	(0.110)***	(0.101)***	(0.102)***
Obs.	1783690	1741733	1057363	1920359	1860232	1135953
Mean dep. var.	13.289	13.292	13.182	13.370	13.369	13.271
Mean dep. var. (NT)	13.253	13.272	13.161	13.935	13.963	13.919
R^2	0.001	0.023	0.049	0.009	0.045	0.073
Log No. Exams	No	Yes	No	No	Yes	No
School FE	No	Yes	No	No	Yes	No
School-Exam FE	No	No	Yes	No	No	Yes

Notes: Dependent variable is the school-level grade of each student in each exam in each year. Columns 2 and 5 include a control for the (log of the) number of exams taken in each school in each year, as well as school fixed effects; columns 3 and 6 control for school-subject fixed effects (only the 5 subjects identified in Table 2 are considered in this column). Robust standard errors, allowing for clustering at the school level. Significance levels: *: 0.1; **: 0.05; ***: 0.01.

Table B4: **Effects on external grades, year by year**

Treated group:	Continent public schools			Continent public schools		
Control group:	Azores and Madeira schools			Private schools		
	(1)	(2)	(3)	(4)	(5)	(6)
2002	0.581	0.647	0.419	0.309	0.401	0.164
	(0.169)***	(0.184)***	(0.233)*	(0.138)**	(0.105)***	(0.130)
2003	0.553	0.554	0.279	0.371	0.501	0.162
	(0.153)***	(0.163)***	(0.126)**	(0.103)***	(0.076)***	(0.114)
2004	0.112	0.028	0.046	0.184	0.166	0.363
	(0.092)	(0.100)	(0.134)	(0.111)*	(0.100)*	(0.135)***
2005	0.601	0.571	0.329	0.743	0.801	0.826
	(0.101)***	(0.106)***	(0.136)**	(0.101)***	(0.092)***	(0.116)***
2007	0.111	0.024	0.395	0.171	0.034	0.433
	(0.094)	(0.098)	(0.111)***	(0.090)*	(0.098)	(0.115)***
2008	1.114	1.079	1.779	1.386	1.197	1.777
	(0.192)***	(0.199)***	(0.231)***	(0.110)***	(0.122)***	(0.148)***
2009	0.973	0.993	1.433	1.335	1.156	1.497
	(0.173)***	(0.177)***	(0.202)***	(0.122)***	(0.131)***	(0.154)***
2010	0.608	0.639	1.154	1.250	1.042	1.515
	(0.194)***	(0.206)***	(0.243)***	(0.122)***	(0.138)***	(0.163)***
2011	0.580	0.574	0.958	1.228	1.023	1.486
	(0.189)***	(0.205)***	(0.211)***	(0.172)***	(0.193)***	(0.208)***
Treated	0.946			-0.326		
	(0.150)***			(0.154)**		
Treated-2002	-0.025	0.005	0.130	0.248	0.259	0.386
	(0.173)	(0.188)	(0.237)	(0.143)*	(0.112)**	(0.138)***
Treated-2003	-0.053	-0.002	0.135	0.129	0.055	0.252
	(0.157)	(0.167)	(0.133)	(0.109)	(0.084)	(0.121)**
Treated-2004	0.043	0.094	0.308	-0.028	-0.062	-0.009
	(0.099)	(0.105)	(0.140)**	(0.117)	(0.104)	(0.141)
Treated-2005	0.116	0.111	0.373	-0.026	-0.131	-0.123
	(0.106)	(0.109)	(0.142)***	(0.106)	(0.097)	(0.123)
Treated-2007	-0.241	-0.285	-0.310	-0.301	-0.318	-0.348
	(0.099)**	(0.100)***	(0.115)***	(0.095)***	(0.102)***	(0.118)***
Treated-2008	-0.307	-0.407	-0.548	-0.579	-0.546	-0.546
	(0.195)	(0.199)**	(0.234)**	(0.114)***	(0.126)***	(0.152)***
Treated-2009	-0.296	-0.418	-0.587	-0.658	-0.595	-0.650
	(0.176)*	(0.178)**	(0.205)***	(0.127)***	(0.136)***	(0.158)***
Treated-2010	-0.212	-0.342	-0.455	-0.853	-0.761	-0.816
	(0.197)	(0.206)*	(0.246)*	(0.127)***	(0.142)***	(0.167)***
Treated-2011	-0.416	-0.521	-0.582	-1.064	-0.985	-1.110
	(0.192)**	(0.205)**	(0.214)***	(0.175)***	(0.197)***	(0.211)***
Obs.	1783689	1741732	1057362	1920358	1860231	1135952
Mean dep. var.	10.608	10.621	10.153	10.728	10.738	10.294
Mean dep. var. (NT)	9.827	9.873	9.257	11.260	11.321	10.962
R^2	0.008	0.050	0.102	0.009	0.061	0.114
Log No. Exams	No	Yes	No	No	Yes	No
School FE	No	Yes	No	No	Yes	No
School-Exam FE	No	No	Yes	No	No	Yes

Notes: Dependent variable is the national exam grade of each student in each exam in each year. Columns 2 and 5 include a control for the (log of the) number of exams taken in each school in each year, as well as school fixed effects; columns 3 and 6 control for school-subject fixed effects (only the 5 subjects identified in Table 2 are considered in this column). Robust standard errors, allowing for clustering at the school level. Significance levels: *: 0.1; **: 0.05; ***: 0.01.

Table B5: **Effects on (standardized) external grades, year by year**

Treated group:	Continent public schools			Continent public schools		
Control group:	Azores and Madeira schools			Private schools		
	(1)	(2)	(3)	(4)	(5)	(6)
2002	0.040	0.062	0.022	-0.068	-0.054	-0.087
	(0.045)	(0.048)	(0.059)	(0.038)*	(0.029)*	(0.032)***
2003	0.051	0.057	0.019	-0.041	-0.018	-0.060
	(0.044)	(0.047)	(0.032)	(0.028)	(0.020)	(0.025)**
2004	0.026	0.005	-0.032	-0.020	-0.035	-0.006
	(0.023)	(0.026)	(0.034)	(0.029)	(0.025)	(0.031)
2005	0.011	0.005	-0.026	-0.019	-0.013	0.001
	(0.029)	(0.030)	(0.032)	(0.027)	(0.024)	(0.027)
2007	0.039	0.017	0.066	0.047	0.000	0.041
	(0.030)	(0.030)	(0.030)**	(0.023)**	(0.025)	(0.030)
2008	0.065	0.054	0.133	0.073	0.015	0.065
	(0.062)	(0.064)	(0.064)**	(0.029)**	(0.031)	(0.039)*
2009	0.051	0.062	0.117	0.133	0.081	0.122
	(0.054)	(0.054)	(0.052)**	(0.032)***	(0.034)**	(0.041)***
2010	0.045	0.058	0.106	0.170	0.104	0.143
	(0.066)	(0.069)	(0.069)	(0.033)***	(0.037)***	(0.043)***
2011	0.110	0.113	0.164	0.216	0.155	0.196
	(0.061)*	(0.065)*	(0.060)***	(0.046)***	(0.050)***	(0.055)***
Treated	0.263			-0.117		
	(0.042)***			(0.044)***		
Treated-2002	-0.043	-0.038	-0.002	0.077	0.092	0.121
	(0.046)	(0.049)	(0.060)	(0.039)**	(0.031)***	(0.034)***
Treated-2003	-0.055	-0.046	-0.000	0.047	0.039	0.089
	(0.045)	(0.048)	(0.034)	(0.030)	(0.022)*	(0.027)***
Treated-2004	-0.028	-0.020	0.050	0.022	0.020	0.024
	(0.025)	(0.027)	(0.036)	(0.030)	(0.025)	(0.033)
Treated-2005	-0.011	-0.016	0.036	0.021	0.001	0.006
	(0.030)	(0.031)	(0.034)	(0.029)	(0.026)	(0.029)
Treated-2007	-0.041	-0.057	-0.074	-0.053	-0.051	-0.052
	(0.031)	(0.030)*	(0.031)**	(0.025)**	(0.026)*	(0.032)
Treated-2008	-0.070	-0.096	-0.147	-0.085	-0.070	-0.083
	(0.062)	(0.063)	(0.065)**	(0.030)***	(0.032)**	(0.040)**
Treated-2009	-0.055	-0.095	-0.129	-0.154	-0.133	-0.149
	(0.055)	(0.053)*	(0.053)**	(0.034)***	(0.036)***	(0.042)***
Treated-2010	-0.048	-0.089	-0.119	-0.194	-0.160	-0.175
	(0.066)	(0.069)	(0.069)*	(0.035)***	(0.038)***	(0.044)***
Treated-2011	-0.117	-0.150	-0.183	-0.249	-0.221	-0.242
	(0.062)*	(0.065)**	(0.060)***	(0.047)***	(0.051)***	(0.055)***
Obs.	1783689	1741732	1057362	1920358	1860231	1135952
Mean dep. var. (NT)	-0.206	-0.193	-0.232	0.145	0.162	0.181
R^2	0.003	0.046	0.072	0.004	0.057	0.086
Log No. Exams	No	Yes	No	No	Yes	No
School FE	No	Yes	No	No	Yes	No
School-Exam FE	No	No	Yes	No	No	Yes

Notes: Dependent variable is the standardized national exam grade of each student in each exam in each year. Columns 2 and 5 include a control for the (log of the) number of exams taken in each school in each year, as well as school fixed effects; columns 3 and 6 control for school-subject fixed effects (only the 5 subjects identified in Table 2 are considered in this column). Robust standard errors, allowing for clustering at the school level. Significance levels: *: 0.1; **: 0.05; ***: 0.01.

Table B6: **Effects on grade inflation, year by year**

Treated group:	Continent public schools			Continent public schools		
Control group:	Azores and Madeira schools			Private schools		
	(1)	(2)	(3)	(4)	(5)	(6)
2002	-0.770	-0.837	-0.572	-0.693	-0.736	-0.482
	(0.183)***	(0.184)***	(0.233)**	(0.111)***	(0.107)***	(0.129)***
2003	-0.816	-0.818	-0.526	-0.684	-0.781	-0.422
	(0.162)***	(0.164)***	(0.149)***	(0.089)***	(0.079)***	(0.101)***
2004	-0.283	-0.161	-0.098	-0.492	-0.440	-0.630
	(0.098)***	(0.106)	(0.150)	(0.081)***	(0.082)***	(0.100)***
2005	-0.815	-0.743	-0.448	-1.068	-1.062	-1.113
	(0.121)***	(0.119)***	(0.140)***	(0.076)***	(0.082)***	(0.106)***
2007	-0.103	0.019	-0.321	-0.169	-0.029	-0.359
	(0.086)	(0.086)	(0.112)***	(0.084)**	(0.085)	(0.091)***
2008	-1.270	-1.164	-1.823	-1.337	-1.133	-1.665
	(0.160)***	(0.162)***	(0.193)***	(0.108)***	(0.116)***	(0.134)***
2009	-1.079	-1.014	-1.450	-1.097	-0.928	-1.229
	(0.126)***	(0.118)***	(0.171)***	(0.111)***	(0.112)***	(0.122)***
2010	-0.713	-0.642	-1.070	-0.880	-0.707	-1.102
	(0.133)***	(0.129)***	(0.189)***	(0.107)***	(0.115)***	(0.127)***
2011	-0.649	-0.541	-0.892	-0.762	-0.611	-1.007
	(0.125)***	(0.135)***	(0.163)***	(0.121)***	(0.139)***	(0.151)***
Treated	-0.937			-0.232		
	(0.141)***			(0.136)*		
Treated-2002	-0.000	-0.011	-0.079	-0.077	-0.109	-0.170
	(0.186)	(0.187)	(0.237)	(0.117)	(0.112)	(0.137)
Treated-2003	0.120	0.082	0.015	-0.012	0.047	-0.089
	(0.165)	(0.167)	(0.154)	(0.095)	(0.085)	(0.108)
Treated-2004	-0.079	-0.124	-0.356	0.131	0.150	0.177
	(0.104)	(0.109)	(0.155)**	(0.088)	(0.088)*	(0.107)
Treated-2005	-0.105	-0.119	-0.362	0.148	0.197	0.303
	(0.125)	(0.122)	(0.145)**	(0.082)*	(0.087)**	(0.113)***
Treated-2007	0.095	0.132	0.162	0.162	0.172	0.200
	(0.090)	(0.089)	(0.115)	(0.088)*	(0.087)**	(0.096)**
Treated-2008	0.321	0.364	0.495	0.389	0.326	0.337
	(0.162)**	(0.163)**	(0.196)**	(0.112)***	(0.118)***	(0.137)**
Treated-2009	0.333	0.376	0.572	0.352	0.286	0.351
	(0.130)**	(0.120)***	(0.174)***	(0.115)***	(0.115)**	(0.126)***
Treated-2010	0.272	0.313	0.400	0.438	0.372	0.432
	(0.137)**	(0.132)**	(0.192)**	(0.112)***	(0.118)***	(0.131)***
Treated-2011	0.486	0.499	0.579	0.599	0.565	0.694
	(0.129)***	(0.137)***	(0.168)***	(0.126)***	(0.143)***	(0.156)***
Obs.	1783689	1741732	1057362	1920358	1860231	1135952
Mean dep. var.	2.500	2.489	2.852	2.459	2.448	2.798
Mean dep. var. (NT)	3.265	3.238	3.746	2.490	2.455	2.775
R^2	0.016	0.056	0.155	0.012	0.058	0.155
Log No. Exams	No	Yes	No	No	Yes	No
School FE	No	Yes	No	No	Yes	No
School-Exam FE	No	No	Yes	No	No	Yes

Notes: Dependent variable is the difference between the internal (school) grade and the external (national exam) grade of each student in each exam in each year. Columns 2 and 5 include a control for the (log of the) number of exams taken in each school in each year, as well as school fixed effects; columns 3 and 6 control for school-subject fixed effects (only the 5 subjects identified in Table 2 are considered in this column). Robust standard errors, allowing for clustering at the school level. Significance levels: *: 0.1; **: 0.05; ***: 0.01.

Table B7: **Robustness: 'intensity' of competition for promotions**

Dependent variable:	External grade			Difference between external and internal grade		
	(1)	(2)	(3)	(4)	(5)	(6)
Share of senior teachers						
After	0.114	0.124	0.708	-0.087	-0.057	-0.715
	(0.091)	(0.095)	(0.109)***	(0.082)	(0.083)	(0.095)***
Share	2.839			-2.221		
	(0.244)***			(0.196)***		
Share-After	-0.425	-0.573	-0.889	0.613	0.676	1.048
	(0.173)**	(0.173)***	(0.209)***	(0.147)***	(0.145)***	(0.177)***
Mean share	0.542	0.543	0.538	0.542	0.543	0.538
R^2	0.013	0.045	0.087	0.013	0.044	0.131
Diff. between senior and 'titular' shares						
After	-0.024	-0.091	0.350	0.147	0.213	-0.266
	(0.076)	(0.073)	(0.085)***	(0.068)**	(0.064)***	(0.080)***
Share diff.	1.491			-1.624		
	(0.572)***			(0.464)***		
(Share diff.)-After	-0.476	-0.506	-0.619	0.499	0.510	0.588
	(0.387)	(0.369)	(0.432)	(0.339)	(0.320)	(0.397)
Mean share diff.	0.194	0.195	0.193	0.194	0.195	0.193
R^2	0.001	0.045	0.087	0.003	0.044	0.130
Obs.	1459572	1445781	866530	1459572	1445781	866530
Mean dep. var.	10.712	10.716	10.283	2.383	2.378	2.707
Log No. Exams	No	Yes	No	No	Yes	No
School FE	No	Yes	No	No	Yes	No
School-Exam FE	No	No	Yes	No	No	Yes

Notes: Dependent variable is the national exam grade of each student in each exam in each year (left panel) or the difference between the internal and external grades (right panel). Data only includes students in public schools. Top panel: 'Share [of senior teachers]' is the proportion of tenured teachers in student i's school whose career stage in school year 2006/07 would have allowed them to apply for 'titular' status (i.e., the upper pay scale following the reform). Bottom panel: 'Share diff.' is the difference between the share of 'senior' tenured teachers (as defined above) and the proportion of 'titular' teachers. Columns 2 and 5 include school fixed effects and a control for the (log of the) number of exams taken in each school in each year; columns 3 and 6 control for school-subject fixed effects (the 5 main subjects are considered in this column: Portuguese, Maths, History, Biology & Geology, and Physics & Chemistry). Robust standard errors, allowing for clustering at the school level. Significance levels: *: 0.10; **: 0.05; ***: 0.01.

Table B8: **Robustness: municipality-year-level economic controls**

Dependent variable:	External grade			Difference between external and internal grade		
	(1)	(2)	(3)	(4)	(5)	(6)
A. Islands control group						
After	-0.151	0.252	0.778	0.235	-0.083	-0.576
	(0.166)	(0.147)*	(0.177)***	(0.126)*	(0.128)	(0.154)***
Continent	0.817			-0.828		
	(0.141)***			(0.113)***		
Continent-After	-0.355	-0.500	-0.751	0.307	0.385	0.630
	(0.141)**	(0.129)***	(0.157)***	(0.109)***	(0.115)***	(0.142)***
Obs.	1565668	1528802	915243	1565668	1528802	915243
Mean dep. var.	10.665	10.678	10.236	2.427	2.417	2.754
Mean dep. var. (NT)	9.900	9.945	9.318	3.192	3.162	3.680
R^2	0.015	0.046	0.090	0.018	0.046	0.137
B. Private control group						
After	0.405	0.454	0.902	0.032	-0.019	-0.349
	(0.151)***	(0.155)***	(0.211)***	(0.130)	(0.108)	(0.146)**
Public	0.059			-0.393		
	(0.136)			(0.130)***		
Public-After	-0.916	-0.708	-0.879	0.431	0.303	0.400
	(0.131)***	(0.137)***	(0.187)***	(0.100)***	(0.087)***	(0.122)***
Obs.	1695839	1641335	989488	1695839	1641335	989488
Mean dep. var.	10.785	10.794	10.378	2.388	2.377	2.701
Mean dep. var. (NT)	11.345	11.402	11.081	2.399	2.366	2.654
R^2	0.020	0.057	0.103	0.014	0.048	0.137
Economic controls	Yes	Yes	Yes	Yes	Yes	Yes
Log No. Exams	No	Yes	No	No	Yes	No
School FE	No	Yes	No	No	Yes	No
School-Exam FE	No	No	Yes	No	No	Yes

Notes: Dependent variable is the national exam grade of each student in each exam in each year (left panel) or the difference between the internal and external grades (right panel). Benchmark data and specifications but including extra control variables (log wages, female ratio, average years of schooling, and log workforce size per *concelho*). Columns 2 and 5 include school fixed effects and a control for the (log of the) number of exams taken in each school in each year; columns 3 and 6 control for school-subject fixed effects (the 5 main subjects are considered in this column: Portuguese, Maths, History, Biology & Geology, and Physics & Chemistry). Robust standard errors, allowing for clustering at the school level. Significance levels: *: 0.10; **: 0.05; ***: 0.01.