Manager recruitment at scale: upheaval, sorting, and demotion in the public sector*

Renata Lemos† Caio Piza‡ Rogerio Santarrosa§ Daniela Scur¶

This draft: September 29, 2026

Abstract

Peru's 2014 school principal selection reform required all incumbents to pass a centralized exam or be replaced, bundling disruption from managerial change with the prospect of higher-quality replacements. Using inverse-probability-weighted difference-in-differences, we estimate that treated schools' achievement fell 0.12 standard deviations in the first post-reform year, recovering by 2018. The disruption dominated in low-amenity, rural, and small schools. While incoming principals at these schools scored higher than the incumbents they replaced, the improvement was smaller than at high-amenity schools—and not large enough to offset the disruption cost. Schools absorbing demoted ex-principals as teachers experienced an additional first-year decline that attenuated in subsequent years, showing how at-scale meritocratic reform can compound disadvantage at the schools it was designed to help.

*We thank participants at seminars and conferences at Oxford University, Queen's University, George Mason University, Wilfrid Laurier University, INSPER, SREE Conference, CAF Research Department, and the World Bank. We also thank Jose Mola, Felipe Macedo Dias, and Lucia Valdivieso for excellent research assistance.

†World Bank. CEPR; CEP-LSE; IGC. E-mail: rlemos@worldbank.org.

‡World Bank. E-mail: caiopiza@worldbank.org.

§Insper. E-mail: rogeriobs2@insper.edu.br.

¶Monash University. CEPR, CEP-LSE. Email: dscur@monash.edu.

1 Introduction

Identifying, selecting and hiring managers is challenging, and doing this well matters: better managers lead to better productivity and outcomes throughout the economy (Bertrand and Schoar; 2003; Fenizia; 2022; Kala; 2024). Establishments in the public sector, in particular, often face a greater challenge in recruiting the best managers as wages tend to be lower and more rigid (Baig et al.; 2021), and teams include a mix of meritocratic and discretionary appointments (Besley et al.; 2022). Education systems account for a large share of public establishments and school principals – the managers of schools – are only seldom appointed by public competition. In Latin America, historically only about one-third of principals were appointed as such (Figure A1), though an increased focus on building meritocratic systems has started to shift this pattern.

Evidence on how meritocratic recruitment affects the selection of public-sector managers mostly comes from gradual reforms. This allows for an empirical analysis comparing establishments that were "treated" earlier versus later but, also, crucially allows for adjustment time within the broader system. For example, the introduction of competitive hiring of school principals and hospital CEOs in Chile led to gains in establishment performance, *over time* (Muñoz and Otero; 2025; Muñoz and Prem; 2024). These gradual reforms often allow time for lower quality managers to exit the system via retirement or voluntary separation, while also allowing time for higher quality candidates to enter public service. When reform is sudden and large-scale, the adjustment instead rests on the existing pool of public-sector workers, affecting both the composition of managers who remain in leadership positions and the allocation of displaced managers within the system.

In this paper we study the impact of a rapid change in the nomination structure of public school principals in Peru towards a large-scale, public civil service exam that affected over 13,000 schools simultaneously. All principals were required to take a special one-off exam (*Excepcional* exam), and those who failed would be replaced. The new principals were to be drawn from a national pool of candidates who passed the regular qualifying exam (*Acceso*) normally available to all existing senior teachers and school principals, with newly expanded minimum qualifying requirements. Those who were fired as principals were offered teaching positions, sometimes in the same school and sometimes within the same system in a different school. The result was a "great reshuffle" of school managers across the country.

We first document the large upheaval in school leadership following the implementation of the reform, and estimate the causal impact of principal change on student outcomes. Only about one-third of the incumbents who sat the *Excepcional* exam passed, with the

vast majority losing their posts to higher-scoring newcomers. Using a matched difference-in-differences design, we estimate that achievement significantly *fell* by approximately 0.12 standard deviations in the first post-reform year in schools that lost their incumbent principal, only reverting to the mean over the next 3 to 4 years. This is the opposite direction of the impact expected following the introduction of a better manager. We also find substantial heterogeneity in this impact: the negative effects are driven by "low amenity" areas, rural and small schools. "High amenity" and urban schools have small and insignificant effects, consistent with receiving strong enough replacements to offset disruption costs.

To characterize this finding we propose a conceptual framework organized around four channels. First, meritocratic selection should, in principle, raise average principal quality and improve outcomes (Colonnelli et al.; 2020; Xu; 2018). But in this setting three other forces potentially at play that could push the other way: assortative matching, disruptive upheaval, and disruptive demotions. The second factor, assortative matching, comes from the national competition (Acesso) allowing higher-scoring candidates to choose their placement. "Best" candidates thus might choose to sort toward more desirable schools, leaving low-amenity areas with a thinner residual pool of replacements. In the public sector especially, this type of sorting can potentially also create a loss of context-specific knowledge (Burks et al.; 2015; Voth and Xu; 2019): the replacement of a local principal with a community "outsider"—even in higher amenity areas—can erase the value of a principal's local knowledge. A third factor is that replacing a leader is disruptive in itself: a sudden, at-scale change in school leadership severs institutional continuity and can negatively impact outcomes despite the replacement's quality (Akhtari et al.; 2022; Busso et al.; 2024). Finally, the closed labor system creates a fourth factor: failed principals were removed from their leadership posts and transferred back into teaching—sometimes in the same schools. These demotions can be disruptive in two ways: they may create frictions between the new principals and the "old guard" and, even absent such frictions, principals might have depreciated teaching skills after several years in leadership positions. Either channel would further negatively affect student outcomes. Thus, whether the reform helps or hurts a given school depends on whether the quality gain from a better principal is large enough to outweigh the disruption and demotion costs that come with the change.

Based on this framework, we then present evidence on these channels. In general, the reform significantly changed the characteristics of principals: incoming principals were younger, scored about 15 points higher on the entrance exam relative to outgoing ones, and were often first-time principals (promoted from teaching positions). Consistent with the sorting channel, higher-amenity schools attracted the strongest replacements, and low-amenity schools

were more likely to receive a principal no better than the one they lost. Thus, where the improvement was smallest, the disruption dominated: the negative effects load entirely onto rural, small, and high-need schools (differential effects of -0.09 to -0.12 standard deviations). Consistent with the hypothesis that *how much better* the principal is matters, we descriptively show that schools gaining a sufficiently higher-scoring principal saw achievement rise within two years and did not experience a first-year decline. Finally, consistent with the demotion channel, the low-amenity schools were most likely to absorb a demoted ex-principal as a classroom teacher, and schools that did so experienced a further decline of about 0.05 standard deviations in the first post-reform year (significant at the 10% level). Taken together, the picture is one of compounding disadvantage: schools in well-positioned areas receive a principal good enough to overcome the upheaval of transition, while schools in disadvantaged areas face the full force of disruption *without* the offsetting benefit of a meaningfully better leader—and may additionally absorb a demoralized former principal as a teacher; at least in the short run.

We contribute to several strands of literature in economics and public policy. First, we contribute to the literature on personnel and public sector productivity (Besley et al.; 2022). Our results highlight the importance of accounting for adjustment costs when moving to large-scale, rapid selection systems, and considering the potential general equilibrium implications of positive assortative matching of such reforms, including sorting and demotion spillovers that are typically absent from studies of more gradual reforms. Second, we contribute to the literature on the importance of personnel practices in the public sector. We show that an at-scale civil service exam can capture managerial skills that are correlated with good school performance, though simply having such skills may not be enough to outweigh other co-existing factors during a period of upheaval. Third, we add to the evidence on civil service reform design. Like the Colombian teacher reform (Busso et al.; 2024) and the US Pendleton Act (Moreira and Pérez; 2024), Peru's experience illustrates that meritocratic selection at the point of entry is necessary but not sufficient: transition management, candidate supply, and organizational spillovers all determine whether the reform achieves its goals.

Our findings also have important policy implications. First, our results show that *how much better* the manager is matters: replacing a manager of unknown or relatively poor quality may not yield positive results if the incoming manager is not a substantial enough improvement to overcome countervailing factors (or, of potentially worse quality). Second, our results suggest that the negative impact was concentrated in "low-amenity", mostly rural and small schools, where the candidate pool is relatively thinner and replacement principals were not meaningfully better than the incumbents they replaced. The negative overall average

reflects this heterogeneity across schools: the within-school disruption cost dominates in low-amenity schools, while in high-amenity schools the cost is more nearly offset (or matched) by the quality improvement. Third, the demotion of principals into teaching roles created a novel and quantitatively important negative spillover; future reforms should consider the downstream effects of what happens to those who do not pass the threshold, particularly when they remain in the system in different roles.

2 Context and data

In this section we outline the historical context of principal selection in Peru, structure of the 2014 reform, and the conceptual framework around what could be the expected outcomes of such a broad, at-scale policy change in this setting, describing the available data throughout.

2.1 Historical context of principal selection in Peru (pre-reform)

Peru's approach to selecting school principals has shifted repeatedly over three decades. In 1984, the *Ley del Profesorado* (LP, "Law for the Teaching Profession"), established a highly decentralized process in which regional or local educational authorities selected principals based on a mix of experience, professional background, and aptitude tests (this was, in fact, the origin of the *Acceso* examination series). A 1990 amendment deepened the hyper-local selection process: authorities prioritized internal candidates (teachers within the school) before opening competition externally.

By 2000, the Ministry of Education (MINEDU) mandated national-level examinations for both teachers and principals, centralizing the process and reducing the weight of subjective criteria. This move towards centralization was met with resistance, and by 2006 there was a return to more decentralized processes with evaluation committees, interviews, and the reinstatement of criteria from the original LP. The last *Acceso* examination under the old regime (the 11th overall) took place in 2005. In 2007, the *Ley de Carrera Pública Magisterial* (LCPM, "Law for the Public Teaching Career") prioritized the inclusion of teachers into a career path of promotion to principal positions, but left many posts under temporary assignments until the 2014 reform. Data from UNESCO's Regional Comparative and Explanatory Studies show that in 2013, the year before the reform, over 55% of principals had been appointed by local authorities or community actors, about 10% through school elections, and only about one-third through civil service examinations (UNESCO-LLECE; 2013). By 2019, the civil-service share had grown to 46% (UNESCO-LLECE; 2019).

2.2 The 2014 reform

In 2012, a new law, the *Ley de Reforma Magisterial* (LRM, "Law of Teaching Reform"), aimed to overhaul the teaching career path and principal selection processes by returning to centralized, meritocratic examinations. Under the new system, all principals in schools with more than five classrooms would need to have passed an examination to hold a post of school principal. Those without a passing score would be offered their prior teaching positions starting the following year.[1] To document these transitions, we use administrative data provided by the Ministry of Education including all candidates who participated in the 2014 *Excepcional* and *Acceso* examinations.[2] The data includes each candidate's school assignment, demographic information (location, education level, gender and age) and examination outcomes. We merge this to the Ministry of Education's administrative allocation dataset (Nexus), which tracks all teacher and principal assignments between 2013 and 2020.[3]

The reform was implemented in two phases. First, facing strong opposition from incumbent principals, the government ran a one-off "exceptional" exam (called *Excepcional*) in August 2014, giving incumbents a chance to retain their positions. All 13,750 principals in eligible schools were summoned to take it. These schools were predominantly regular basic education institutions: over three-quarters were primary (54.75%) and secondary (22.42%) schools.[4] Of these, 64% were the single principal at their primary school, and about 62% took the exam. Only 34% of those then passed it. Those who either did not take the exam or failed were allowed a second chance by sitting the regular *Acceso* examination in November 2014, but they now competed against all other eligible candidates and were not guaranteed their current school even if they passed.[5] Three-quarters of principals who failed *Excepcional* attempted *Acceso*, and 40% of them passed. Among those who did not attempt *Excepcional*, only one quarter sat for *Acceso*, with 36.5% passing. Compliance with the policy was high: principals who failed both exams or refused to take them were mostly removed from their posts (Figure 1).[6]

[1] Figure A2 summarizes the timeline of the reform and the steps.

[2] We also have data from Peru's 2018-2019 *Evaluación del Desempeño en Cargos Directivos* (manager performance evaluation), which we use in Appendix A.2 (Table A2) as a validity check on the principal exam scores: the correlation between a principal's exam score and their 2018 performance evaluation indicates that the exam captures skills predictive of later managerial performance.

[3] Nexus records point-in-time employee snapshots, rather than monthly observations. To account for changes during the school year, such as transfers or delayed appointments, we use two snapshots per year: end of April (start of school year) and end of December (end of school year).

[4] The remaining schools provided early childhood education (18.09%), alternative basic education (2.73%), technical-vocational education (1.32%), or special basic education (0.68%).

[5] Panel (b) of Figure A2 traces the flow of principals through the two exams down to the analysis sample.

[6] A small share of principals sued the government to retain their posts and remained in position tem-

The *Acceso* examination also had an important change in eligibility criteria relative to prior exams. Previously, candidates needed to have reached the "fourth scale" of teaching tenure (approximately 11 years of service) to sit the exam. The 2014 reform allowed teachers in the "first scale" to qualify if they could demonstrate at least one year of prior experience in a managerial role. This expanded eligibility led to a large influx of substantially younger and less experienced candidates, which we confirm in the data later in the analysis.[7]

Most had experience, and specifically in their schools. The average principal had been a principal for over 10 years, and had been in their school (as a principal) just over 6 years. In terms of their time-use, principals dedicate an average of 35.8 hours per week to their school duties, and 40% of principals also serve as teachers at their school. Few have duties outside their school: only about 12% teach at other schools, 6% hold concurrent principal posts and 20% hold other jobs (half of those are educational). Principals seem to take knowledge of managerial tasks seriously: in the prior two years, 64.2% report having completed specialization or postgraduate studies in educational administration or management.

While in general it is not clear whether public sector exams can usefully identify worker quality (c.f. Aman-Rana; 2025; Bertrand et al.; 2020; Dahis et al.; 2025), in this case the evidence suggests the exams do capture skills linked to school performance (see Appendix A.2 for details). Figure 2 shows the distribution of exam scores for incoming and outgoing principals in schools that experienced a leadership change, though with the caveat that we only have this information for half of the sample. As our principal-quality measure we use the Text Comprehension subtest score, taking the *Excepcional* score when available and falling back on *Acceso* otherwise.[8] On average, incoming principals scored higher than those they replaced, consistent with the exam's goal of raising the quality threshold.

2.3 Data

We combine four administrative datasets: Peru's *Evaluación Censal de Estudiantes* (ECE) for student outcomes; the *Censo Escolar* school census and the 2007 national Census for school and district characteristics; the Ministry of Education's (MINEDU) Nexus system for principal and teacher assignments; and MINEDU's records of the 2014 *Excepcional* and

porarily, but were eventually removed.

[7]Specifically, treated schools' incoming principals were on average younger, scored substantially higher on the entry examination, and were more likely to have moved across schools to take the role than the incumbents they replaced (Table 2).

[8]Appendix A.2 details the construction and the reasons we use this subtest rather than the MINEDU composite final score.

Acceso examinations for principal exam scores. From these, we construct three samples: (i) an analysis sample of 2,543 schools whose incumbent principal sat the 2014 *Excepcional* exam, used to estimate the reform's direct effect on student outcomes (intent-to-treat); (ii) a "spillover" sample of 8,839 schools used to study the indirect effect of receiving a demoted ex-principal as a classroom teacher; and (iii) a broader "turnover" sample of 4,165 schools – drawn from the 6,898 schools with classifiable 2014 to 2015 principal transitions – used for the descriptive system-wide analysis of how leadership turnover relates to student achievement. The unit of analysis is the school-year, and we first provide a set of summary statistics for the universe of schools from each data source and outline the specific analysis samples below.

2.3.1 Peruvian administrative datasets

ECE: Learning outcomes. Peru's learning outcomes have been historically low, especially relative to other Latin American countries, but they have been improving over time. Our primary outcome data come from Peru's *Evaluación Censal de Estudiantes* (ECE), a national standardized test in math and reading administered at the end of each academic year in all schools with at least five students in the tested grade. We have school-level average scores for second-grade students from 2009 through 2016 and for fourth-grade students in 2016 and 2018. No ECE was administered in 2017. Table A1 reports average student achievement: on the ECE's 500-centered scale, the average Peruvian primary school scored 526 in second-grade math and 544 in reading in 2014.[9]

We treat the school-level average score as a proxy for school quality, as our interest is in how a school's position in the national distribution changed around the reform. To make scores from different grades comparable on a common interpretive scale, we standardize each grade-subject combination independently against the 2016 ECE universe: a school's score is then the standard deviations from the same-grade national mean in 2016. That is, a score of 0.5 in second-grade math and a score of 0.5 in fourth-grade math both represent half a standard deviation above the same-grade 2016 distribution, so the standardized score has the same interpretive scale across grades. The combined outcome takes the standardized second-grade score in 2009–2015, the average of the standardized second- and fourth-grade scores in 2016 (the bridge year in which both grades are observed), and the standardized fourth-grade score in 2018.[10] We merge student outcomes data with school characteristics from the school

[9]The ECE reports school results as a *medida promedio*: student achievement is estimated with an Item Response Theory (Rasch) model and linearly transformed onto a scale with the ECE 2007 distribution as the base-year (Oficina de Medición de la Calidad de los Aprendizajes (UMC); 2016).

[10]We do this because neither grade has student data far back or forward enough for this analysis, and

census (*Censo Escolar*), including enrollment, school resources, and promotion, dropout, and repetition rates.

Censo Escolar: school characteristics. Peru's primary school system includes approximately 51,400 schools across the country. The 2014 reform included about one-fifth of the country's primary schools, as the majority of primary schools have fewer than 5 classrooms. Table A1 reports summary statistics in 2014. At the time of the reform, the average primary school in Peru had 115 students, about 41% of schools were located in urban areas, and about half of households in the average school's district had basic needs met. Principals were majority male (about 67%), 49 years old on average, and earned PEN 2,402 per month — about twice the average monthly labor income in Peru in 2014 (Instituto Nacional de Estadística e Informática; 2018). Teachers were about 50% female, 44 years old on average, and earned PEN 1,643 per month. Schools had an 86% promotion rate and a 2% dropout rate.

Amenity index. To capture heterogeneity in how desirable a particular school posting is, we construct an *amenity index* as a standardized weighted average of four school-level characteristics: an urban location indicator, log enrollment (log number of 2014 students), a school resource infrastructure index measured in 2014, and the share of households with met basic needs at the district level from the 2007 Census. The index is meant to proxy for the non-wage attributes that make a school more or less attractive to prospective principals. The four components are correlated with one another, so combining them into a single composite captures the underlying desirability of a posting and reduces the multiple-testing burden. Each component is standardized to mean zero and standard deviation one, and the components are then combined into a single index using inverse-covariance weights as in Anderson (2008). We classify schools into high-amenity (above median) and low-amenity (below median) groups for the heterogeneity analysis.[11]

Nexus and Exam Records: principals and teachers. For principal and teacher assignments, we use the Ministry of Education's administrative allocation system (Nexus),

under the assumption that either grade is a good proxy for school-level performance they are both equally useful. The per-grade standardization removes mechanical differences in the scale and difficulty of the second- and fourth-grade tests, so a coefficient on the combined outcome is interpretable as a change in the school's position in the same-grade national distribution. Further, the 2016 bridge year averages the standardized scores from both grades, smoothing the transition between the second-grade-only pre-period and the fourth-grade-only post-period. Finally, we verify in Table A4 that the short-run pattern of the main result holds on a strictly within-grade panel restricted to second-grade scores through 2016.

[11] Figure A4 shows the distribution of the amenity index components.

which tracks all teacher and principal assignments. We use two snapshots for each year: end of April (start of school year) and end of December (end of school year). We merge this with principal examination records from MINEDU that include each candidate's school assignment, demographic information (location, education level, gender, and age), and examination outcomes for the 2014 *Excepcional* and *Acceso* exams. We use the Nexus payroll panel to identify, for each school and year, how principals moved between schools. Further, we can identify whether schools received a demoted principal as a teacher; a key variable we construct below.

Figure 3 summarizes the movements of teachers and principals across schools around the time of the reform. In the two years prior to the reform, when selection systems were decentralized and mostly local, the share of lateral moves of principals between schools was under 1%, and the share of teacher-to-principal promotions was under 5%. In 2015, the year following the reform, more than a third of schools promoted a teacher into the principal role, and roughly one in ten saw a lateral move; this rate stayed between 10 and 20% for promotions in subsequent years. Lateral moves among principals trended upward through 2020, suggesting that principals were aware of and engaged with the movement opportunities afforded by the new centralized system. Demotions spiked to 8.5% in 2015 and stabilized around 2.5% through 2018.

Demotion indicator. We construct the demotion indicator at the individual level in three steps. First, for each job-school position (*plaza*) we flag a turnover whenever the individual occupying that position changes from one year to the next. Second, for each individual whose position changed, we record whether the move was within the same school or to a different school. Third, we classify the type of move by comparing the worker's role (principal or teacher) before and after: a *demotion* is an individual observed as a principal in year $t-1$ and as a tenured teacher (*nombrado*) in year t (we exclude placements into temporary contract posts). The analogous categories for promotions, lateral moves, and new entries are defined symmetrically. We restrict the demotion indicator to individuals who had been summoned to the *Excepcional* examination, so the moves we count are reform-induced rather than ordinary churn.[12]

[12]If a teacher in 2014 had no record or was on a temporary contract, they would not be coded as demoted in 2015. Only schools with more than 8 classes have a full time principal, while smaller schools have a senior teacher appointed to exercise principal duties in a part time capacity, for a financial bonus. In this case, if these teachers lost their part time principal duties they would not be considered demoted in the data from the Nexus database, as they would have full time teacher contracts in both 2014 and 2015.

2.3.2 Analysis samples

Intent-to-Treat sample. Our main *analysis* sample comprises the schools whose incumbent principal sat the 2014 *Excepcional* exam, and contains 2,543 schools. Our "treated" schools are those where the incumbent principal sat the *Excepcional* examination but did *not* secure continued tenure at their own school: they neither passed *Excepcional* nor were placed back at their original school through the subsequent *Acceso* competition, and were therefore assigned to leave their post. Our "control" schools are those where the incumbent retained leadership of the same school, either by passing *Excepcional* or by being placed back at their original school through *Acceso*. A principal placed at a *different* school through *Acceso* is not counted as having retained their post, since the original school still undergoes a leadership change. The vast majority of principals who were assigned to leave their post did, indeed, do so (Figure 1). Relative to the average primary school in the country, our analysis sample schools are larger (239 vs. 115 students), more urban (51% vs. 41%), and located in slightly better-resourced districts (54% vs. 50% of households having met basic needs).

Spillover sample. Our *spillover* sample includes a much larger number of schools (8,827 schools), as demoted principals from the reform were placed as teachers in schools throughout the system; not just in reform-eligible schools. We use this sample to study the indirect effect of the reform on schools that absorbed a demoted ex-principal as a classroom teacher. The sample is drawn from the same IPW-weighted universe as the main ITT analysis. *Treated* schools are those that received a demoted ex-principal (as defined above) as a teacher in 2015, and *control* schools are those that did not. Relative to the average primary school in the country, our spillover sample schools are larger (214 vs. 115 students), similarly urban (42% vs. 41%), and located in comparably-resourced districts (49.7% vs. 50.1% of households having met basic needs).

Broader turnover sample. For the descriptive system-wide analysis in Section 4.4 we construct a broader sample of schools that experienced a principal transition between 2014 and 2015, regardless of whether the transition was directly induced by the reform's exam-failure mechanism. Starting from the full payroll universe of 70,585 unique schools, we restrict to primary and secondary schools (40,365), to schools with exactly one principal in each of 2014 and 2015 (10,899), and to those for which we can recover an evaluation score for both the outgoing and the incoming principal — using the *Excepcional* exam score when available, and the *Acceso* exam score otherwise. The final sample is 6,898 schools

with classifiable transitions, which we partition into two groups according to the relative score of the incoming versus outgoing principal: *Better* (incoming score strictly higher than outgoing) and *No Better* (incoming score equal or lower).[13] Of these, 4,165 schools also have the student-outcome data needed for the analysis.

3 Conceptual framework

The goal of the reform was to improve the quality of principals by ensuring a minimum level of competence and improving selection into the profession. We outline four channels through which the reform could affect student outcomes suggesting the direction of the effect in square brackets. The net effect is, thus, ex-ante ambiguous.

Channel 1: Selection of better principals in general [likely positive]. Rule-based selection through exams can be useful for screening candidates based on objective, measurable skills (Bertrand et al.; 2020; Dahis et al.; 2025), but selection based on discretion can also leverage private information such as candidates' tacit knowledge of particular schools and communities that exams cannot capture (Burks et al.; 2015; Voth and Xu; 2019). The evidence on the selection of bureaucrats through rule-based, meritocratic systems is mixed (Araujo et al.; 2020; Busso et al.; 2024; Estrada; 2019; Moreira and Pérez; 2024; Ornaghi; 2019), though generally tends towards positive. More specifically, the evidence for using meritocratic systems for selection of *managers* is mostly positive, albeit more limited: Muñoz and Prem (2024) and Muñoz and Otero (2025), for example, find competitive hiring of school principals and hospital CEOs in Chile (respectively) led to improvements in establishment performance. In our setting, the data confirms that incoming principals score higher on the dimensions the exam measures (Section 4, Table 2). This channel pushes toward improved outcomes, though such positive impacts likely operate with a lag: principals' effects on student outcomes tend to accumulate over multiple years (Coelli and Green; 2012), and analogous rule-based recruitment reforms elsewhere have shown that the candidate pool continues to shift across successive cycles (Bedoya et al.; 2026). Exams are unlikely to capture all aspects of quality; however, we show in Appendix A.2 that, in this case, the exam did, in fact, capture a signal of quality, albeit noisy.

[13] Mixing scores from *Excepcional* and *Acceso* is approximate, since the two exams differ in scale and in test-taker population; we discuss this in more detail in Appendix A.2.

Channel 2: Sorting across schools [ambiguous]. The reform opened competition nationwide rather than maintaining local preference, so higher-scoring candidates could exercise preferences over where they were placed. This sorting creates winners and losers across schools, and the net direction of the overall effect depends on whether gains and losses across schools net out. First, candidates sort on amenities: more attractive schools (urban, larger, better-resourced) draw the 'best' principals (on observables, at least), while rural and low-amenity schools face a thinner residual pool and often receive replacements who are 'no better' or potentially 'worse' than the incumbents they replace (see also Busso et al.; 2024, on the importance of candidate supply). Second, a centralized system that allows candidates to sort on amenities across regions is likely to leave schools where local and tacit knowledge matters most (typically rural and remote communities) with candidates who not only ranked lower overall but also lack the community-specific knowledge that discretion-based selection might have leveraged. Thus, if replacement candidates are only a small improvement in terms of ability relative to the incumbents they replace (more likely in thinner markets), they may actually be worse if the local knowledge matters in that particular community. Both sorting mechanisms compound disadvantage in already underserved areas, making the aggregate net effect across all schools an empirical question.

Channel 3: Impact from turnover-driven "upheaval" [likely negative]. Replacing any leader, regardless of quality, disrupts institutional continuity, relationships, and management practices. Akhtari et al. (2022) show that bureaucratic upheaval following political turnover in Brazil, which included replacement of school personnel at both managerial and non-managerial levels, led test scores to fall by 0.05 to 0.08 standard deviations. Leaver et al. (2024) extend this finding and show that the negative effect is driven by schools where the principal was replaced, with a significant drop in management quality.[14] Importantly, much of the existing evidence compares establishments that were *already experiencing turnover* but received managers of different quality (Muñoz and Otero; 2025; Muñoz and Prem; 2024). In our setting, the reform created disruption where there otherwise would have been none.

Channel 4: Demotion spillovers [likely negative]. A salient feature of this reform is that "fired" managers were not necessarily removed from the system; they were transferred back to teaching positions, often within the same schools where they had served as principals. This created an unusual disruption channel, as ex-principals potentially saw this as a

[14]See also Bartanen et al. (2019) on principal turnover in the US, and Miller (2013) on the multi-year recovery trajectory after principal changes.

demotion. These demoted principals both may have depreciated teaching skills after years not teaching, (Dinerstein et al.; 2022) as well as negative feelings or low morale regarding their change in status. This could create friction within the school, especially if the incoming principal is younger, less experienced, or not embedded into the local community.

Net effect. The overall impact depends on the relative magnitudes of these four channels and varies across school types. For a school in a high-amenity area, for example, the negative costs may be offset by receiving a substantially better principal through favorable sorting. A school in a low-amenity area, however, faces disruption, may receive a replacement who is not a meaningful improvement, and may additionally absorb a demoralized demoted principal. The reform could thus have heterogeneous effects, improving outcomes in well-positioned schools while compounding disadvantage in already-struggling ones.

4 Effects of large-scale, rapid reform

The scale and timing of the 2014 reform yielded a system-wide reshuffling of principals, as well as a substantial influx of newly promoted principals from teaching positions. This led to both a "forced" upheaval in many schools, and a new changed landscape of school leaders characteristics. In this section we describe our identification strategy and present results on the average impact, heterogeneous effects, demotion spillovers and general turnover patterns.

4.1 Estimating intent-to-treat with propensity score weights

The core of our empirical strategy is to compare schools where principals were removed to a weighted control group of schools that retained their principals — matched on pre-reform student outcomes — under the assumption that both groups would have followed parallel trends absent the reform. We employ a matched difference-in-differences design to estimate the impact of the reform on student outcomes. Our treated schools are those where the incumbent principal sat the *Excepcional* examination but did not secure continued tenure at their own school: they neither passed *Excepcional* nor were placed back at their original school through the subsequent *Acceso* competition, and were therefore assigned to leave their post.[15] Our control schools are those where the incumbent retained leadership of the

[15]For the main analysis, we focus on the sample of schools whose principal took the *Excepcional* exam, which allows us to observe exam scores for both the outgoing and incoming principal. In Appendix A.6 we verify that adding to the treated group the schools whose incumbent principal did not sit the *Excepcional* exam at all — conceptually defensible since the reform's "treatment" was the offer to leave the post regardless

same school, either by passing *Excepcional* or by being placed back at their original school through *Acceso*. A principal placed at a *different* school through *Acceso* is not counted as having retained their post, since the original school still undergoes a leadership change.

As the definition of treatment is based on the offer to retain or leave their position rather than actual take-up, our setup estimates the intent-to-treat (ITT) effect of the reform. Figure 1 and Appendix Figure A2 document that compliance with the reform was high and largely one-sided: of the incumbent principals who passed the *Excepcional* examination, an overwhelming share retained their original-school post, while roughly three-quarters of those who failed lost their post by either failing or not sitting the *Acceso* fallback.[16]

To construct a comparable control group, we follow Abadie (2005) and use inverse-probability weighting (IPW)[17]. We first estimate the probability of being treated conditional on observables using a logit regression of the treatment indicator on a rich vector of school-level covariates:

$$Z_s = \alpha_0 + \alpha_1 \sum_{\tau=1}^{5} X_{s,t-\tau} + u_s, \tag{1}$$

where $X_{s,t-\tau}$ includes lagged standardized math and reading scores (five lags), promotion and dropout rates, enrollment (log number of students), an urban indicator, regional fixed effects, school resource measures, and the district-level share of households with unmet basic needs (NBI). We then construct weights as (Hirano and Imbens; 2001):

$$w_s = \begin{cases} 1 & \text{if } Z_s = 1 \text{ (treated)} \\ \frac{\widehat{p}_s}{1-\widehat{p}_s} & \text{if } Z_s = 0 \text{ (control)} \end{cases} \tag{2}$$

where $\widehat{p}_s$ is the estimated propensity score. We restrict the sample to the region of common

of test-taking behavior — still yields a significant negative effect on student achievement in the first post-reform year, with the effect attenuating in the second year. This is most visible within low-amenity schools, where the first-year decline is larger in absolute magnitude when non-takers are added (-0.19 versus -0.15 standard deviations). Under our conceptual framework, this pattern is consistent with the non-takers having been even more negatively selected ex-ante, since their replacements appear to have been "good enough" to overcome the disruption cost by the second post-reform year.

[16]While some principals sued to retain their posts, the share who did so is small and most eventually left. We focus on ITT rather than the average treatment on the treated (ATT) because the small remaining non-compliance — failed principals who appealed and temporarily retained their posts — dilutes treatment intensity but does not reverse assignment. Under this near one-sided non-compliance, our ITT estimates can be read as a lower bound on the average treatment effect for the principals whose treatment status was determined by the exam outcome.

[17]Unlike Abadie (2005) though, we estimate the propensity score paramatrically and use weights as in Hirano and Imbens; 2001

support.[18] Using these weights, we estimate the following event study specification with 2014 as the base year, on the weighted sample (Abadie; 2005; Blundell and Costa Dias; 2009; Hirano and Imbens; 2001, as in):

$$y_{st} = \alpha + \sum_{\tau \neq 2014} \delta_\tau (Z_s \times \mathbf{1}[t = \tau]) + \mu_s + \tau_t + \varepsilon_{st} \tag{3}$$

where y_{st} is the average student outcome in school s in year t, Z_s is the treatment indicator, $\mathbf{1}[t = \tau]$ are year indicators with the year of the reform examinations (2014) as the omitted category, μ_s are school fixed effects, τ_t are year fixed effects, and ε_{st} is the error term. Standard errors are clustered at the school level. The parameters of interest, δ_τ, estimate the intent-to-treat dynamic effect of the reform on treated schools in each year relative to 2014.[19]

The key identifying assumption is that, after reweighting on observables (estimated propensity score), treated and control schools would have followed parallel trends in student outcomes absent the reform. For the heterogeneity analysis, we both use the same control group and include interaction terms to maintain comparability, but we verify that the coefficients are similar when we re-estimate the control group using the propensity score and IPW weights within each subgroup (e.g., low- vs. high-amenity schools) excluding that variable from the weighting function to ensure covariate balance within subgroups.[20]

As outlined in our conceptual framework, it is not clear ex-ante what the impact of a change in principal would be in the setting of the Peruvian reform. While in theory incoming principals should be better than outgoing principals, for there to be a positive impact relative to schools that retained a principal who passed, the incoming principals would have to represent an improvement that would outweigh the impact of the managerial turnover disruption. Among treated schools where we can compare incoming and outgoing principal scores, about 20% ended up with an incoming principal whose score was no higher

[18] Appendix materials include the distributions of propensity scores for the weighted and unweighted samples, confirming adequate overlap.

[19] A potential concern with our two-way fixed effects specification is the bias documented by Goodman-Bacon (2021) and de Chaisemartin and D'Haultfœuille (2020) when treatment timing varies across units. In our setting, the recent concern with specification bias (as in de Chaisemartin and D'Haultfœuille; 2020; Goodman-Bacon; 2021) is mitigated because treatment is sharp and simultaneous: all treated schools experienced the principal change in the same year (2015) as a result of the 2014 examinations. See Bertrand et al. (2004) on inference in DiD designs with serial correlation, and Angrist and Pischke (2009) for a general discussion.

[20] We include the interaction specification instead of the sub-group specific weights version to be able to compare coefficients. However, the coefficients of regressions using sub-group specific weights are not meaningfully different. In Appendix A.4 we report this analysis.

than the outgoing one's (Panel B of Figure 2); the majority had a "better" principal as the average quality distribution shifted to the right (Panel A).

4.2 Overall impact on student outcomes

Figure 4 presents the event-study estimates of the reform's impact on average student achievement. Overall, Peru had substantial learning gains in the period since 2012, but relative to the control schools, treated schools performed 0.1 standard deviations worse in the short term. There are no significant differences between treated and control schools in the pre-period, but in the first year after the reform (2015), treated schools show a statistically significant decline of approximately 0.116 standard deviations on average. By the second year, the negative effect attenuate to about 0.079 standard deviations, marginally significant at the 10% level and the gap closes to (not significant) 0.026 standard deviations by 2018.

Table 1 reports the DiD coefficients for the average test scores as well as math and reading separately, alongside promotion and dropout rates. We find no impact on either promotion or dropout rates, suggesting that the lower achievement is not driven by differential changes in school progression patterns or in the composition of students continuing through the school. We discuss the composition of the workforce, principal and school characteristics below.

While the average short term impact is negative, there is substantial heterogeneity in terms of where the impact is most pronounced. Our conceptual framework proposes that the net impact of the reform depends on the relative strength of the positive selection channel versus the ambiguous impact of sorting and the negative impact of disruption and spillover channels. Indeed, Figure 4b shows that the negative effect is driven by schools that score below the median in our amenity index.

4.3 Mechanisms and heterogeneity: sorting on amenities

The composition of students stayed relatively stable after the reform (at least in terms of dropouts and promotions), and so did other school- and teacher-level composition. Specifically, the share of new teachers, average teacher age, average teacher salary, and the share of female teachers are not significantly different across amenity type. Principal characteristics, however, did significantly change.

To show this, we estimate the triple-interaction IPW-weighted difference-in-differences

specification used throughout this section:

$$y_{st} = \alpha + \beta \cdot (Z_s \times \text{Post}_t) + \delta \cdot (Z_s \times \text{Post}_t \times G_s) + \gamma \cdot Z_s + \lambda \cdot G_s + \phi_t + \varepsilon_{st}, \tag{4}$$

where y_{st} is the characteristic of interest for school s in year t, Z_s is the treatment indicator, $\text{Post}_t = \mathbf{1}$ for years after 2015, G_s is a binary indicator for the focal subgroup (e.g. high or low amenity), and ϕ_t are year fixed effects. Observations are weighted by the school-level IPW weights w_s from equation (2), and standard errors are clustered at the school level. The coefficient β captures the ITT×Post effect for the reference group ($G_s = 0$); the triple-interaction coefficient δ captures the *differential* ITT×Post effect for the focal group ($G_s = 1$). Because all subgroups share the same sample and average IPW weights, the differential coefficients δ are comparable. While in this framework the control group may not have optimal weights, we verify that re-estimating optimal weights for the control group for each iteration yields similar results.[21]

For school-level outcomes (whether the school was given a demoted principal, share of new teachers) and average teacher characteristics, which vary year by year, we estimate equation 4 on the 2009–2018 panel with the outcome of interest as the dependent variable in place of student achievement. For incoming principal characteristics, where the variable is fixed at the incumbent's 2014 value or the replacement's 2015 value, we restrict the panel to those two years and estimate a difference-in-differences specification on the two-year balanced panel:

$$y_{st} = \alpha + \gamma Z_s + \psi \, \text{Post}_t + \beta \, (Z_s \times \text{Post}_t) + \varepsilon_{st}, \tag{5}$$

where y_{st} is the characteristic of the principal at school s in year t, Z_s is the treatment indicator, and Post_t is the post-reform indicator. Observations are weighted by the school-level IPW weights w_s from equation 2, standard errors are clustered at the school level, and the sample is restricted to the balanced panel of schools where the characteristic is observed in both years. The coefficient β measures how much the incoming principal differs from the outgoing incumbent at treated schools relative to the corresponding change at control schools.

[21] Because our analysis sample is reweighted using inverse-probability weights to balance treated and control schools on pre-reform observables, a raw comparison of treated and control means would ignore those weights and produce a biased contrast. In Appendix A.4 we verify that re-estimating each subgroup effect on its own subgroup-specific IPW-reweighted sample (so that within-subgroup balance is optimized at the cost of comparability across subgroups) yields qualitatively similar magnitudes, suggesting the cross-subgroup comparability gained by using a common sample with average weights does not come at a meaningful cost to within-subgroup balance (Table A11).

We re-estimate the specification separately within each focal subgroup with subgroup-specific IPW weights, and report the resulting β coefficient in each cell of Table 2.[22]

Table 2 reports how the reform changed the type of principals that schools received, overall and separately for low- and high-amenity schools (Panel A). On average, incoming principals were about 4.3 years younger than the within-school change at control schools and scored about 18 points higher on the *Acceso* exam (both significant at the 1% level). But these gains were far from uniform across the amenity distribution: high-amenity schools received principals who were 5.3 years younger and scored 20 points higher, versus 3.3 years younger and 15 points higher at low-amenity schools (differences statistically significant at the 1% level). Incoming salaries and the share female rose at high-amenity schools but not at low-amenity ones. This is the sorting pattern at the center of our framework: the strongest replacements flowed to the most desirable schools.

Turning to student outcomes, Figure 5 shows the relationship between principals' *Acceso* exam scores and the school's amenity index, separately for principals who left their post (leavers) and those who replaced them (newcomers). Two patterns stand out. First, higher-amenity schools tend to have both leavers and newcomers who scored higher, suggesting that better-scoring principals sorted into schools with more desirable working conditions on both sides of the reform. Second, newcomers score uniformly higher than leavers across the amenity distribution, consistent with a positive selection effect on average, and the gap varies with amenity. The improvement in principal quality from the reform is thus concentrated in higher-amenity schools, while low-amenity schools experience the same disruption with a smaller quality dividend.

This has downstream implications for student outcomes consistent with the results above. If the sorting channel is strong, we would expect to see more negative effects in low-amenity schools, where the quality of incoming principals is not meaningfully better than the outgoing ones, and thus the disruption is not offset by a quality improvement. In contrast, in high-amenity schools, the improvement in principal quality may be sufficient to offset the disruption, leading to smaller negative effects, null or even positive effects. To unpack this, we estimate equation 4 for the average standardized math and reading score outcome on the analysis sample.

Table 3 reports the heterogeneous post-reform effects. Column (1) reports the average ITT×Post effect on the full IPW-weighted sample. Columns (2) to (8) each interact the baseline with one subgroup dummy. The ITT×Post row reports the effect for the omitted

[22]The underlying ITT×Post coefficients by subgroup, along with standard errors and within-cell sample sizes, are reported in Appendix Tables A8, A9, and A10.

reference group, and each of the Subgroup × ITT × Post rows give the differential for the focal group. In all, there isn't a single factor in the amenity index that drives the full relationship; the negative effects are similar across rural, small, and high needs schools, where the interaction terms add −0.086 to −0.119 standard deviations to the generally small and insignificant reference group.[23] We interpret this as evidence that the sorting of principals into "high amenity schools" (broadly construed) is a key factor in the initial decline in student achievement, suggesting that the upheaval of leadership change could not be overcome in the low amenity schools by principals who were not meaningfully better than the outgoing ones.

4.4 Beyond the ITT: a descriptive view of the broader turnover sample

The reform-induced reshuffle was much wider than the ITT sample: Figure 3 showed that lateral moves among principals and demotions of failed principals into teaching positions both spiked at the reform and stayed elevated for several years afterward — which means that many schools outside our ITT analysis sample also experienced a leadership transition in this window, either because their incumbent moved laterally, was demoted, or because the broader *Acceso* process delivered a new principal under the post-reform rules. To get a system-wide descriptive picture of how this reshuffle reshaped student achievement, we turn to the broader turnover sample defined in Section 2.3.

We estimate an event-study specification on this broader sample, analogous to equation 3 but replacing the ITT indicator with a turnover indicator T_s that equals one for schools whose 2015 principal differs from the 2014 incumbent:

$$y_{st} = \alpha + \sum_{\tau \neq 2014} \delta_\tau (T_s \times \mathbf{1}[t = \tau]) + \mu_s + \tau_t + \varepsilon_{st}, \tag{6}$$

where y_{st} is the school-level average test score in year t, μ_s are school fixed effects, τ_t are year fixed effects; and 2014 (the year of examinations) is the omitted base year. Observations are weighted by inverse-probability weights constructed on the broader-turnover sample's covariates, and standard errors are clustered at the school level. We estimate the specification on the full broader-turnover sample (4,165 schools) and separately on the Better (2,432) and No Better (1,472) subsamples defined in Section 2.3, using subgroup-specific IPW weights in

[23] Appendix Table A11, which estimates each subgroup effect on a subgroup-specific reweighted sample, reports magnitudes within about 0.02 standard deviations of the ones in this table, suggesting that the cross-subgroup comparability gained by using a common sample with average weights does not come at a meaningful cost to within-subgroup balance.

the splits.

Figure 7 Panel (a) reports an event-study analog on this broader sample for schools that received a principal who scored higher on the test ("better") or a principal who scored equal or worse ("no better") on the test, compared to schools that did not have any change in principals.[24] The patterns here are closer to what the literature in this topic has shown: a positive trend following the switch to a better leader. Consistent with the framework where upheaval and disruption have a negative impact, the first year of a "better" principal still yields a null result in terms of student achievement, but after the upheaval subsides the positive impacts begin to show. Specifically, by the second post-reform year (2016), schools that experienced a transition to a better-scoring principal show a positive effect of 0.096 standard deviations on math (significant at the 5% level) and 0.079 standard deviations on the math/reading average (significant at the 10% level). Those that changed to a principal who is no better had a negative but insignificant change in the first year, which also reverts to the mean. This pattern is consistent with our framework's prediction that positive selection can outweigh the disruption cost when the quality improvement is large enough.

Still, this analysis is at best suggestive. The reform changed both who left and who was available to replace them: it pushed out failed principals, centralized the *Acceso* process, and expanded eligibility to first-scale teachers with managerial experience. The bundle of changes preclude a clean separation of the principal-quality channel from the broader changes the reform set in motion. Third, the "no turnover" group serves as our comparison group, as is the conventional approach in the school-leadership literature, but those schools are a selected subset that escaped the reform's pressure rather than a quasi-randomly assigned counterfactual. Thus, we see this analysis as simply descriptive evidence that the system-wide reshuffle reshaped achievement in ways that line up with our conceptual framework's predictions — not as causal estimates of the effect of getting a better principal.

4.5 Spillover impact of demotions

A distinctive feature of Peru's reform is that principals who failed the examinations were not removed from the education system, but rather were moved to classroom teaching positions often in the same schools or nearby schools; a move that was seen by many of the ex-principals as a demotion. This creates a spillover channel: the receiving schools absorb teachers whose teaching skills may have depreciated during their years in management, who may be demor-

[24] Appendix Table A14 reports the underlying coefficients.

alized by the demotion, and thus set up an environment with potential frictions.[25] These demotion spillovers compound the direct effects of the reform. Low-amenity schools, which already face disruption from principal turnover and weaker replacement candidates, are also more likely to absorb demoted principals (Table 2).

Panel (a) of Figure 6 quantifies the demotion spillover using an event-study specification analogous to equation 3, on the spillover sample (schools that did and did not receive a demoted ex-principal as a teacher).[26] In the first post-reform year (2015), schools that received a demoted ex-principal experienced a drop of about 0.05 standard deviations in student achievement (significant at the 10% level for math, reading, and the average), with the effect attenuating but persisting in 2016 and 2018 at roughly -0.03 to -0.05 standard deviations (no longer significant). There is no systematic relationship with dropout or promotion rates. This pattern is consistent with a one-time disruption from absorbing a demoted ex-principal, partially recovering over subsequent years.

Which schools receive a demoted ex-principal reflects administrative and geographic decisions that are not strictly random, so this setup is not free of selection issues and should be interpreted with caution. Still, the fact that the spillover sample is drawn from the same IPW-reweighted universe as the main ITT analysis, and that the event-study specification includes school fixed effects, helps mitigate some of these concerns by controlling for time-invariant differences between schools and by balancing on pre-reform observables. The results suggest that the demotion spillover could be a meaningful channel through which the reform impacted student achievement, particularly in the first year after the reform, when the shock of absorbing a demoted ex-principal was most visible.

5 Discussion and conclusion

Peru's 2014 principal examination reform was ambitious in scope and meritocratic in design. It aimed to raise the quality of school leadership by requiring all principals to sit a centralized examination, replacing those who did not measure up to the stated exam standards. But implementing such a change at scale is complex, and while there was fairly high compliance, the resulting impact on student outcomes was not overall positive – at least in the short run.

Our key insight is that in a system with heterogeneous school quality and finite candidate supply, a nationwide meritocratic competition can backfire and exacerbate inequality in outcomes. Higher-scoring principals exercise their preferences and sort into more desirable

[25] Appendix Figure A3 traces the flows of demoted principals into their new teaching positions.

[26] Appendix Table **??** reports the underlying coefficients.

schools, while schools at the bottom of the amenity distribution receive the residual candidates. When this sorting is combined with the disruption of sudden, at-scale reform and the organizational friction of forced demotions, the result is a widening of the gap between well-served and underserved schools.

We highlight that it is important to consider not just the potential change in principal quality measured by an exam, but also local conditions, sorting incentives and disruption effects. We document that, despite high compliance, student achievement dropped on average. We propose a conceptual framework that rationalizes this finding: while an incoming principal might be ostensibly "better" by scoring higher on the civil service exam, change is inherently disruptive and the improvement in principal quality has to be large enough to make up for the negative impacts. Further, change at scale is complex and needs to account for shifting incentives: if principals are now allowed to choose schools in different districts, better principals might self-sort into better schools. Without embedding compensating differentials in the system, this aggravates the issue in poorer districts as they contend with both the disruption of upheaval and a negative selection in incoming principals. Finally, the question of what happens to displaced school leaders is often ignored: if principals are part of a public sector that does not lead to dismissal from the system, this inherently builds in a spillover effect into the teaching profession where outgoing principals – who may be feeling dejected – are now returning as teachers with depreciated skills.

We propose that the average negative impact we see in the short run, and indeed that many papers in this literature report, likely reflect a combination of these impacts. This interpretation is supported by the evidence: schools that are likely to experience positive selection (high amenity schools) see no significant difference in student performance in the short run, despite having the best incoming principals. While it is possible the principals who scored higher are, in fact, no better in reality, given existing evidence on the negative impact of upheavals it is more likely that the incoming principals were able to at least make up the negative impact of the disruption and the spillover effects of demoted principals if those existed. However, in areas where principals were not necessarily *that* much better, they were not able to overcome those negative effects and the overall impact was negative.

The demotion channel falls along the same lines: low-amenity schools are also more likely to receive a demoted ex-principal as a classroom teacher (Table 2), so the schools that tend to get the weakest replacement quality are also the ones absorbing this additional source of disruption.

Peru's experience is consistent with a recurring pattern across countries that have implemented meritocratic selection reforms at scale. Colombia's 2005 teacher hiring reform

replaced experienced contract teachers with higher-scoring but novice exam-passers, and student outcomes fell (Busso et al.; 2024). The US Pendleton Act of 1883 improved federal employee credentials but did not improve cost-effectiveness, partly because organizations adapted by shifting hiring to exempt positions (Moreira and Pérez; 2024). Across these settings, the common finding is that short-term disruption effects are front-loaded while quality gains materialize slowly, creating a window of negative net impact that can persist for one to several years. When this happens in the education sector, where learning compounds over years, such changes can affect several cohorts of students and hinder their lifelong learning ability and future career prospects.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators, *The Review of Economic Studies* **72**(1): 1–19.

Akhtari, M., Moreira, D. and Trucco, L. (2022). Political turnover, bureaucratic turnover, and the quality of public services, *American Economic Review* **112**(2): 442–493.

Aman-Rana, S. (2025). Meritocracy in a bureaucracy, *Journal of Development Economics* **175**: 103428.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects, *Journal of the American Statistical Association* **103**(484): 1481–1495.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Princeton, NJ.

Araujo, M. D., Heineck, G. and Cruz-Aguayo, Y. (2020). Does test-based teacher recruitment work in the developing world? experimental evidence from ecuador, *Technical report*, IZA Discussion Papers.

Baig, F. A., Han, X., Hasnain, Z. and Rogger, D. (2021). Introducing the worldwide bureaucracy indicators: A new global dataset on public sector employment and compensation, *Public Administration Review* **81**(3): 564–571.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/puar.13355*

Bartanen, B., Grissom, J. A. and Rogers, L. K. (2019). The impacts of principal turnover, *Educational Evaluation and Policy Analysis* **41**(3): 350–374.

Bedoya, J., de Hoyos, R. and Estrada, R. (2026). Rule-based civil service: Evidence from a nationwide teacher reform in Mexico, *Economics of Education Review* **111**: 102771.

Bertrand, M., Burgess, R., Chawla, A. and Xu, G. (2020). The glittering prizes: Career incentives and bureaucrat performance, *The Review of Economic Studies* **87**(2): 626–655.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?, *The Quarterly Journal of Economics* **119**(1): 249–275.

Bertrand, M. and Schoar, A. (2003). Managing with style: The effect of managers on firm policies, *The Quarterly Journal of Economics* **118**(4): 1169–1208.
URL: *https://ideas.repec.org/a/oup/qjecon/v118y2003i4p1169-1208..html*

Besley, T., Burgess, R., Khan, A. and Xu, G. (2022). Bureaucracy and development, *Annual Review of Economics* **14**: 397–424.

Blundell, R. and Costa Dias, M. (2009). Alternative approaches to evaluation in empirical microeconomics, *Journal of Human Resources* **44**(3): 565–640.

Burks, S. V., Cowgill, B., Hoffman, M. and Housman, M. (2015). The value of hiring through employee referrals, *The Quarterly Journal of Economics* **130**(2): 805–839.

Busso, M., Montaño, S., Muñoz-Morales, J. and Pope, N. (2024). The unintended consequences of merit-based teacher selection: Evidence from a large-scale reform in colombia, *Journal of Public Economics* **239**: 105238.

Coelli, M. and Green, D. A. (2012). Leadership effects: School principals and student outcomes, *Economics of Education Review* **31**(1): 92–109.

Colonnelli, E., Prem, M. and Teso, E. (2020). Patronage and selection in public sector organizations, *American Economic Review* **110**(10): 3071–3099.

Dahis, R., Schiavon, L. and Scot, T. (2025). Selecting top bureaucrats: Admission exams and performance in brazil, *Review of Economics and Statistics* **107**(2): 408–425.

de Chaisemartin, C. and D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects, *American Economic Review* **110**(9): 2964–2996.

Dinerstein, M., Megalokonomou, R. and Yannelis, C. (2022). Human capital depreciation and returns to experience, *American Economic Review* **112**(11): 3725–62.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.20201571*

Estrada, R. (2019). Rules versus discretion in public service: Teacher hiring in mexico, *Journal of Labor Economics* **37**(2): 545–579.
URL: *https://www.journals.uchicago.edu/doi/abs/10.1086/700192*

Fenizia, A. (2022). Managers and productivity in the public sector, *Econometrica* **90**(3): 1063–1084.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA19244*

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing, *Journal of Econometrics* **225**(2): 254–277.

Hirano, K. and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization, *Health Services and Outcomes Research Methodology* **2**(3–4): 259–278.

Instituto Nacional de Estadística e Informática (2018). Perú: Indicadores de Empleo e Ingreso por Departamento, 2007-2017, *Technical report*, Instituto Nacional de Estadística e Informática, Lima, Perú.

Kala, N. (2024). The impacts of managerial autonomy on firm outcomes, *Econometrica* **92**(6): 1777–1800.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA19872*

Leaver, C., Lemos, R. and Scur, D. (2024). Understanding School Management with Public Data: A New Measurement Approach and Applications, *CEPR Discussion Paper DP14069 (v.2)*, Centre for Economic Policy Research, London. First published 21 October 2019; this revision 28 July 2024.

Miller, A. (2013). Principal turnover and student achievement, *Economics of Education Review* **36**: 60–72.

Moreira, D. and Pérez, S. (2024). Civil service exams and organizational performance: Evidence from the pendleton act, *American Economic Journal: Applied Economics* **16**(3): 250–291.

Muñoz, P. and Otero, C. (2025). Managers and public hospital performance, *American Economic Review* **115**(11): 4040–4074.

Muñoz, P. and Prem, M. (2024). Managers' productivity and recruitment in the public sector, *American Economic Journal: Economic Policy* **16**(4): 223–253.

Oficina de Medición de la Calidad de los Aprendizajes (UMC) (2016). Reporte técnico de la Evaluación Censal de Estudiantes (ECE 2015), *Technical report*, Ministerio de Educación del Perú, Lima, Perú.
URL: *http://umc.minedu.gob.pe/wp-content/uploads/2016/07/Reporte-Tecnico-ECE-2015.pdf*

Ornaghi, A. (2019). Civil service reforms: Evidence from us police departments, *Mimeo* .

UNESCO-LLECE (2013). Tercer estudio regional comparativo y explicativo (terce) [dataset].

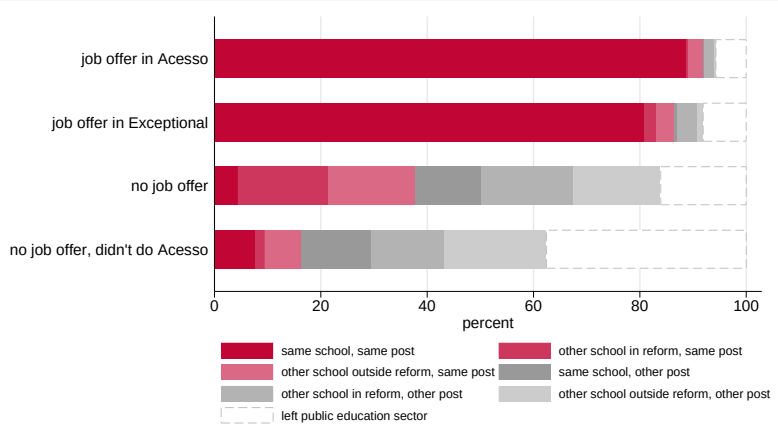
UNESCO-LLECE (2019). Estudio regional comparativo y explicativo (erce) 2019 [dataset].

Voth, H.-J. and Xu, G. (2019). Patronage for productivity: Selection and performance in the age of sail.

Xu, G. (2018). The costs of patronage: Evidence from the british empire, *American Economic Review* **108**(11): 3170–3198.

Tables and Figures

Figure 1: Compliance with Peru's 2014 reform: principals' post-reform status by exam outcome



Note: Data from Nexus administrative data and MINEDU exam records. Post-reform employment status of incumbent Peruvian principals by their 2014 exam outcome. Rows group principals by their offer status: those who received a job offer through *Acceso*, those who received an offer through *Excepcional*, those who failed both exams ("no job offer"), and those who did not sit *Acceso* ("no job offer, didn't do Acceso"). The horizontal bars show the share of principals in each post-reform state in 2015: still serving as principal at the same school and same post (dark red), at a different school but same post (lighter reds), moved to a different post such as classroom teaching (grays), or left the public education sector (dashed outline). Compliance is high: nearly all principals with a job offer remained at the same school in the same role, while most without an offer transitioned to other posts or exited the system.

Table 1: ITT Effects on Student Outcomes, overall and by school amenity

	Test Scores (std. dev.)			Rates	
	Math	Reading	Average	Promotion	Dropout
Overall					
ITT × Post	-0.065***	-0.065**	-0.064**	0.003	-0.001
	(0.025)	(0.026)	(0.025)	(0.003)	(0.001)
Schools	2,543	2,543	2,543	2,543	2,543
School-Year Obs	22,887	22,887	22,887	25,419	25,419
R-squared	0.150	0.137	0.153	0.119	0.194
Mean Dep. Var (2016)	0.162	0.238	0.086	0.912	0.010
Low Amenity Index					
ITT × Post	-0.100**	-0.100**	-0.101**	0.002	0.002
	(0.044)	(0.047)	(0.043)	(0.005)	(0.003)
Schools	1,194	1,194	1,194	1,194	1,194
School-Year Obs	10,746	10,746	10,746	11,930	11,930
R-squared	0.133	0.106	0.150	0.146	0.208
Mean Dep. Var (2016)	-0.185	-0.075	-0.295	0.919	0.010
High Amenity Index					
ITT × Post	-0.005	-0.009	-0.001	0.004	-0.001
	(0.031)	(0.032)	(0.031)	(0.004)	(0.001)
Schools	1,194	1,194	1,194	1,194	1,194
School-Year Obs	10,746	10,746	10,746	11,939	11,939
R-squared	0.219	0.217	0.211	0.086	0.196
Mean Dep. Var (2016)	0.507	0.548	0.466	0.905	0.010

Note: Data from ECE student-outcome records and MINEDU exam records. IPW-weighted difference-in-differences estimates of post-reform ITT effects on student outcomes. Each panel reports the ITT×Post coefficient and standard error (in parentheses) for five outcomes: standardized math, reading, and the math+reading average score (columns 1–3), and school-level promotion and dropout rates (columns 4–5). The "Overall" panel reports the full analysis sample (2,543 schools); the "Low Amenity Index" and "High Amenity Index" panels split the analysis sample at the median of the school amenity index (1,194 schools each). Standard errors clustered at the school level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 2: Principal, school and teacher composition, by school amenity

	All	Low Amenity	High Amenity	p-value (Low=High)
Panel A. Principal characteristics				
Age	-4.312***	-3.312***	-5.276***	0.001
	(0.294)	(0.404)	(0.450)	
Log salary	0.042***	-0.018	0.100***	0.000
	(0.014)	(0.024)	(0.021)	
Female	0.037*	-0.013	0.111***	0.002
	(0.019)	(0.026)	(0.030)	
Test score	17.778***	14.983***	20.423***	0.000
	(0.507)	(0.722)	(0.753)	
Schools	1959	883	960	
Panel B. School and teacher characteristics				
Received demoted outsider	-0.011***	-0.017***	-0.007	0.243
	(0.003)	(0.006)	(0.005)	
% new teachers	-0.001	-0.001	-0.001	0.965
	(0.001)	(0.002)	(0.001)	
Average teacher age	0.295	0.134	0.220	0.834
	(0.210)	(0.345)	(0.216)	
Average teacher salary	0.005	0.000	0.007	0.584
	(0.007)	(0.011)	(0.006)	
Share female teachers	-0.007	-0.007	-0.003	0.796
	(0.009)	(0.015)	(0.007)	
Schools	2543	1194	1194	
Panel C. Incoming principal *(share of treated schools)*				
Moved from origin	0.702***	0.721***	0.648***	0.023
	(0.014)	(0.020)	(0.025)	
Had first-time incoming principal	0.578***	0.589***	0.554***	0.312
	(0.016)	(0.023)	(0.026)	
Had no-better incoming principal	0.075***	0.083***	0.061***	0.194
	(0.008)	(0.011)	(0.012)	
Received demoted insider	0.031***	0.034***	0.026***	0.062
	(0.002)	(0.003)	(0.003)	
Schools	2543	1194	1194	

Per-panel school counts (clusters); Ns vary by cell within a panel, see Table A10 for complete details.

Note: Data from ECE student-outcome records, MINEDU exam records, and Nexus administrative data. Each cell reports the ITT×Post coefficient (standard error in parentheses) from a separate IPW-weighted difference-in-differences regression of the listed characteristic, estimated within all schools and, separately, within low- and high-amenity schools (split at the median of the school amenity index). The final column reports the p-value for the test that the low- and high-amenity coefficients are equal. Principal characteristics are estimated on the balanced 2014–2015 panel; demotion and teacher characteristics on the full 2009–2018 panel. In Panel B, "Received demoted outsider" indicates a school that received, as a teacher in 2015, an ex-principal demoted from a *different* school. Panel C reports shares of treated schools (control schools have no incoming principal): "Moved from origin" is the share whose replacement moved in from another school, "Had first-time incoming principal" the share whose replacement had not previously been a principal, "Had no-better incoming principal" the share whose replacement scored no higher than the outgoing one, and "Received demoted insider" the share whose *own* former principal was demoted to a teaching post at the same school. Detailed coefficients for the full set of subgroups are in Appendix Tables A8, A9, and A10. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 3: Heterogeneous effects: subgroup interactions with ITT×Post

	Outcome variable: student test scores						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ITT × Post	-0.065***	-0.002	-0.007	0.003	-0.016	-0.041	-0.060
	(0.025)	(0.026)	(0.025)	(0.027)	(0.030)	(0.035)	(0.037)
Low Amenity × ITT × Post		-0.109**					
		(0.047)					
Amenity Index Components							
Rural × ITT × Post			-0.105**				
			(0.047)				
Small × ITT × Post				-0.119**			
				(0.047)			
Low Met Needs × ITT × Post					-0.086*		
					(0.048)		
Low Resources × ITT × Post						-0.046	
						(0.049)	
Principal Status							
First-Time × ITT × Post							-0.019
							(0.059)
Schools	2,543	2,543	2,543	2,543	2,543	2,543	2,543
School-Year Obs	22,887	22,887	22,887	22,887	22,887	22,887	22,887
R-squared	0.150	0.240	0.241	0.219	0.247	0.151	0.160
Mean Dep. Var (2016)	0.162	0.162	0.162	0.162	0.162	0.162	0.162

Note: Data from ECE student-outcome records, MINEDU exam records, and Nexus administrative data. Each column reports a separate IPW-weighted DiD regression of equation 4 on the common analysis sample of 2,543 schools with average IPW weights. Column (1) is the baseline with no interaction. Columns (2)–(7) add a subgroup dummy and its interaction with ITT×Post: the ITT×Post row reports the effect for the omitted reference group, and the subgroup-interaction row reports the differential for the focal group. Outcome is the average standardized 2nd- and 4th-grade math and reading score. Robust standard errors clustered at the school level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

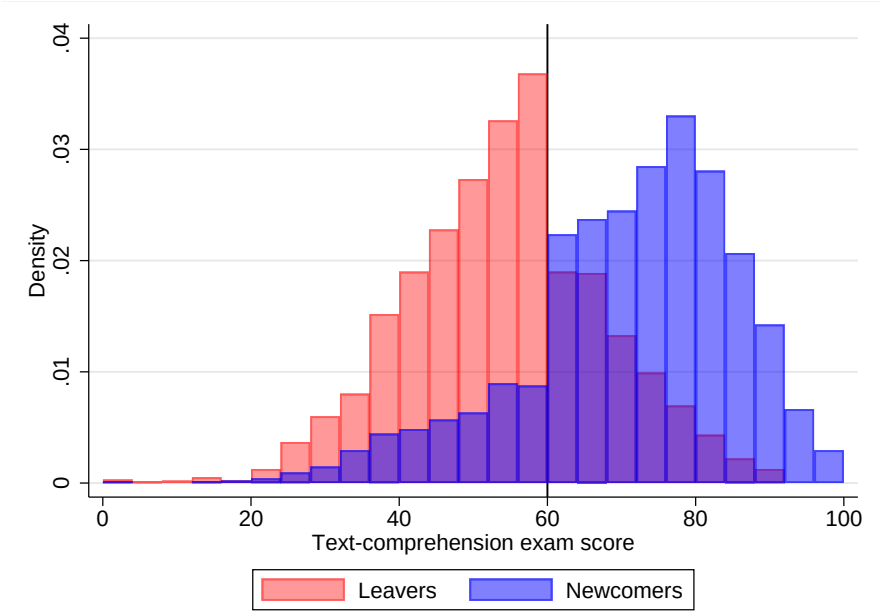
Table 4: Student achievement by demotion type and setting

	Test scores (SD)			Rates	
	Math	Reading	Avg	Promotion	Dropout
Panel A: Insider demotion (reform, principal changed)					
Control group: reform schools with no principal change nor newly demoted teachers					
Group × Post	-0.067**	-0.055*	-0.061*	-0.004	0.000
	(0.034)	(0.031)	(0.031)	(0.004)	(0.002)
Treatment N	396	396	396	396	396
Control N	999	999	999	999	999
Post-reform first year effect	-0.079	-0.054	-0.067	0.000	-0.001
Panel B: Outsider demotion (reform)					
Control group: reform schools with no principal change nor newly demoted teachers					
Group × Post	-0.062**	-0.047*	-0.055*	-0.002	0.000
	(0.031)	(0.029)	(0.029)	(0.004)	(0.002)
Treatment N	478	478	478	478	478
Control N	1053	1053	1053	1053	1053
Post-reform first year effect	-0.116**	-0.115**	-0.116**	0.000	0.002*
Panel C: Outsider demotion (non-reform)					
Control group: non-reform schools with no newly demoted teachers					
Group × Post	-0.041	-0.050*	-0.045	0.001	0.000
	(0.031)	(0.027)	(0.028)	(0.003)	(0.001)
Treatment N	437	437	437	437	437
Control N	4240	4240	4240	4240	4240
Post-reform first year effect	-0.115**	-0.119**	-0.117**	-0.001	0.001
Panel D: Turnover only (reform, principal changed)					
Control group: reform schools with no principal change nor newly demoted teachers					
Group × Post	0.021	0.013	0.017	0.000	0.000
	(0.021)	(0.020)	(0.020)	(0.003)	(0.001)
Treatment N	2678	2678	2678	2678	2678
Control N	1083	1083	1083	1083	1083
Post-reform first year effect	-0.053	-0.067*	-0.060*	0.003	0.000
School FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes

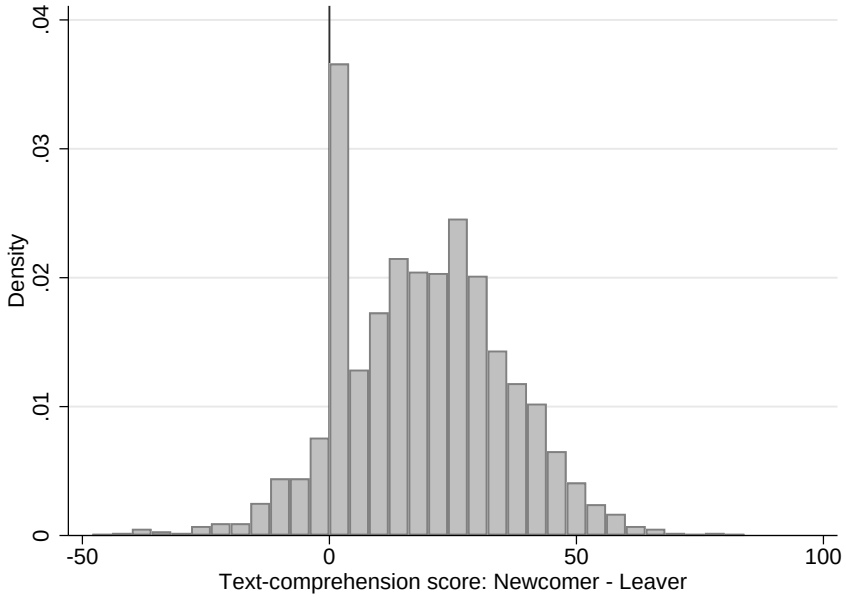
Each cell is the Group×Post coefficient from a difference-in-differences of the panel group against its control group, with school and year fixed effects, inverse-probability weights, and standard errors clustered by school. Panels A, B, and D (reform, summoned schools) use a group-specific IPW against the reform pure control; Panel C (non-reform) uses the any-demotion IPW against non-reform no-demotion controls, pooling (largely unobserved) principal-change status. Panel B pools reform outsider demotions whether or not the school's own principal also changed; the split appears in Appendix Table A13. Non-reform demotions are essentially all outsider. The italicized *Post-reform first year effect* is the 2015 event-study coefficient (the group×2015 term from `i.treat##ib2014.year`, with 2014 omitted). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure 2: Distribution of principal exam scores

(a) Newcomers and leavers distributions, system level

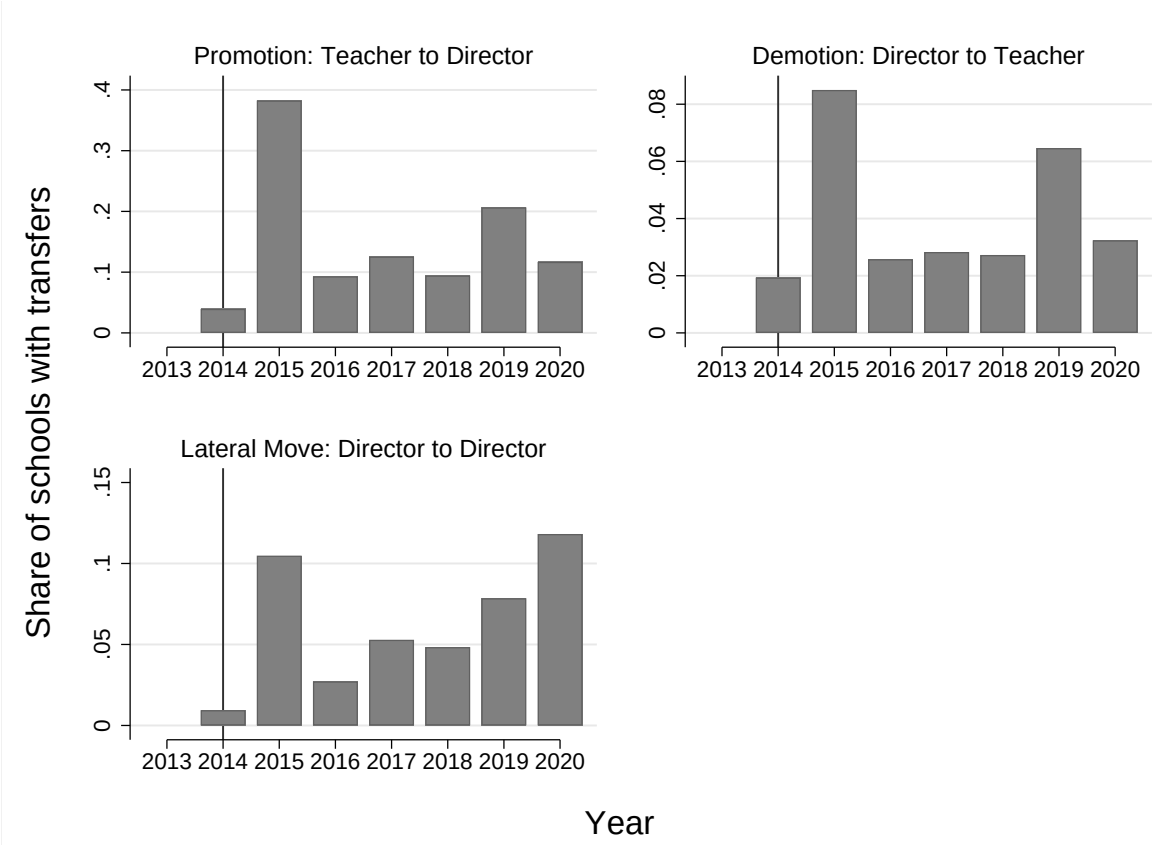


(b) Difference in scores, school level



Note: Data from MINEDU exam records. This figure shows the distribution of exam scores for incoming ("newcomer") and outgoing ("leaver") principals in schools that experienced a change in leadership following the 2014 reform. Panel (a) plots the system-level distributions of newcomer and leaver scores; Panel (b) plots the school-level difference in scores between the incoming and outgoing principal.

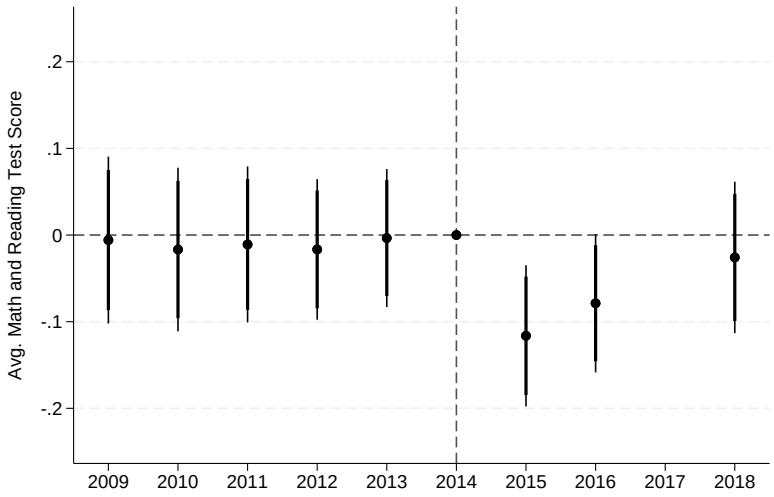
Figure 3: School-level transfers over time



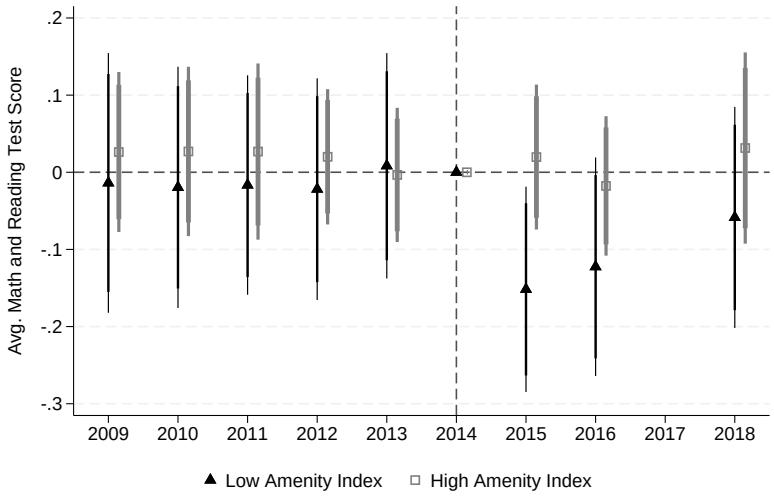
Note: Data from Nexus administrative data, MINEDU. This figure shows the share of schools experiencing different types of principal transitions over time: promotions (teacher to principal), demotions (principal to teacher), and lateral moves (principal to principal). The vertical line marks the year of the reform examinations (2014).

Figure 4: Effect of principal replacement on student achievement: event study

(a) All treated schools

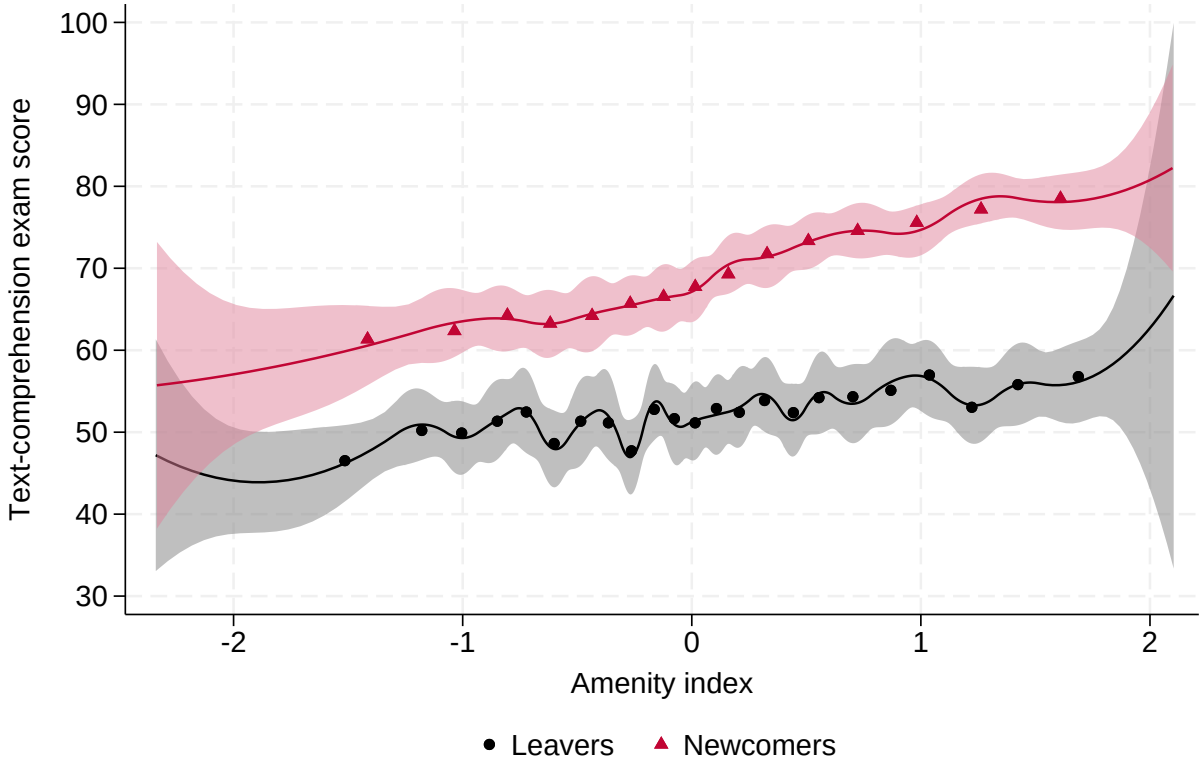


(b) By amenity index: low vs. high



Note: Data from ECE student-outcome records and MINEDU exam records. Event-study estimates of the effect of principal replacement on average student achievement (math and reading combined). The outcome combines 2nd- and 4th-grade standardized scores as described in Section 2.3: 2nd-grade scores for 2009–2015, the average of 2nd- and 4th-grade scores for 2016, and 4th-grade scores for 2018, all standardized to the 2016 universe mean and standard deviation. Treated schools are those where the principal failed the examinations and were supposed to be replaced. Control schools retained their principal. The specification includes school and year fixed effects, with observations weighted using inverse-probability weights. Standard errors are clustered at the school level. The omitted year is 2014 (the year of examinations), marked by the vertical line. Panel (a) pools all treated schools. Panel (b) splits the treated group by the school amenity index at the median, plotting the low-amenity and high-amenity event-study trajectories on the same axes. The underlying annual coefficients are reported in Appendix Tables A5 (overall), A6 (low amenity), and A7 (high amenity).

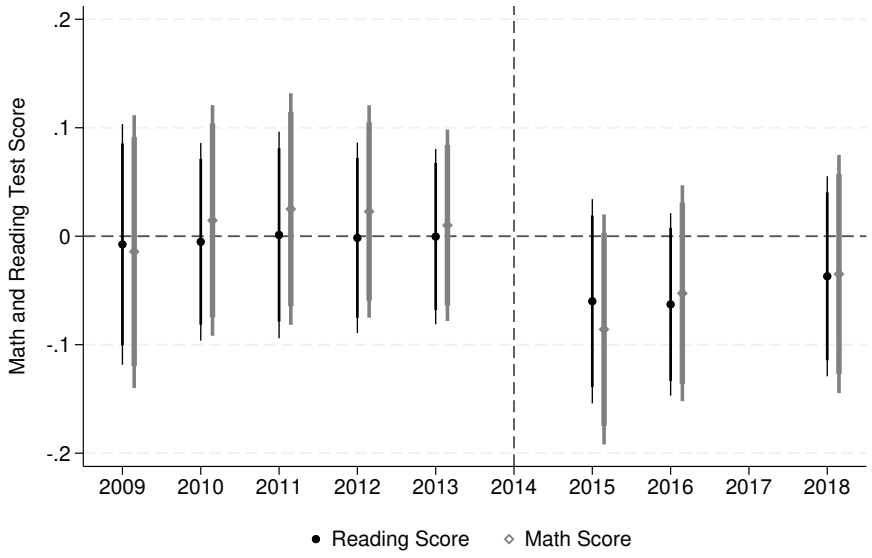
Figure 5: Principal *Acceso* scores by school amenity index: leavers and newcomers



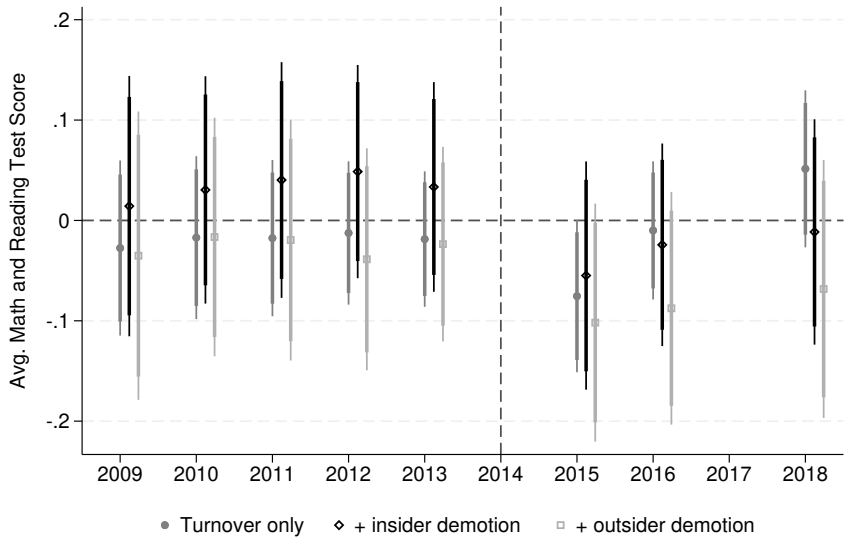
Note: Data from MINEDU exam records (Acceso scores); Nexus administrative data; underlying amenity-index sources described in Section 2.3. Binned scatterplot of the *Acceso* examination score against the school amenity index, separately for outgoing principals (leavers) and their replacements (newcomers) in treated schools. Each marker is the mean score within a bin of the amenity index; the lines are linear fits through the underlying school-level observations. Higher values of the amenity index correspond to schools in more desirable locations (urban, higher resources, fewer households with unmet basic needs). The plot uses Stata's `binsreg` command for the binned scatter and the fitted lines. The y-axis is the Text Comprehension subtest score from the 2014 examinations (*Excepcional* where available, *Acceso* otherwise); see Section 2.3 for details on the score.

Figure 6: Demotion effects in the reform sample: event study

(a) Demotion spillover



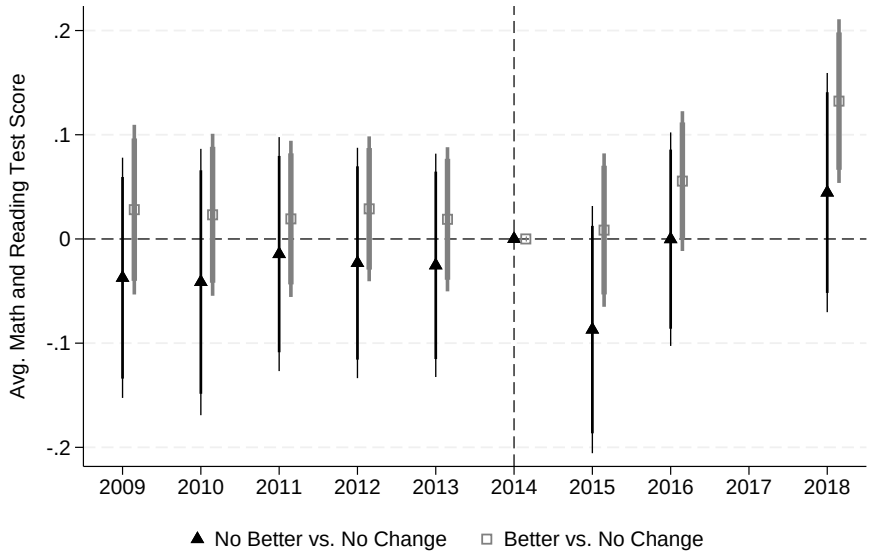
(b) By demotion type: turnover only, insider, outsider



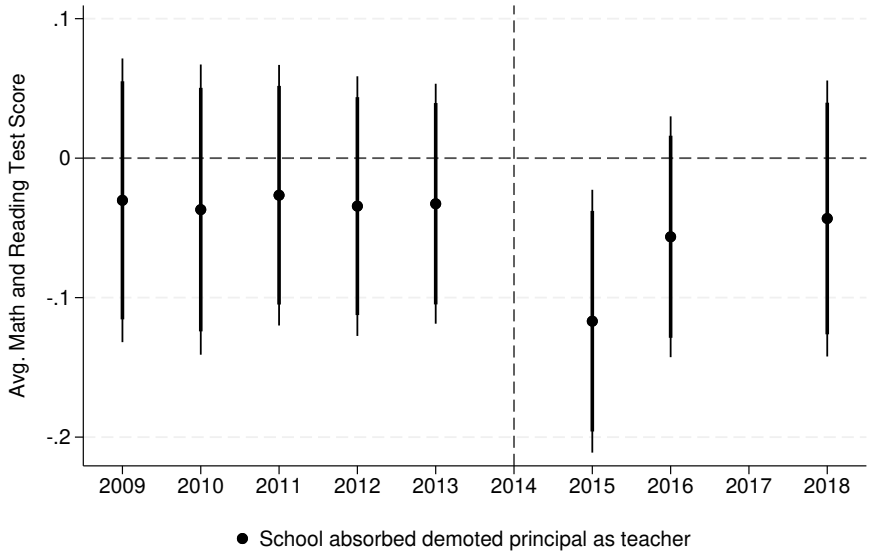
Note: Data from ECE student-outcome records, MINEDU exam records, and Nexus administrative data. Reform (summoned) sample. Panel (a): event-study estimate of the spillover effect of receiving a demoted ex-principal as a teacher in 2015 (spillover sample; standardized math and reading scores as two series). Panel (b): the same effect decomposed against the pure control (schools with no principal change and no demoted teacher; Table 4) — turnover only (principal changed, no demotion), plus insider demotion, plus outsider demotion (combined) — each weighted by a group-specific inverse-probability weight; average standardized math and reading score. The omitted year is 2014; 2017 is a gap (scores not collected). Standard errors clustered at the school level; 90 and 95% confidence intervals shown. Underlying coefficients in Appendix Table **??**.

Figure 7: Spillovers: principal turnover and non-reform demotions

(a) Broader turnover spillovers



(b) Demotion spillover, non-reform schools



Note: Data from ECE student-outcome records, MINEDU exam records, and Nexus administrative data. Panel (a): event-study estimates on the broader turnover sample (Section 2.3), plotting schools that received a better-scoring principal ("Better vs. No Change") and an equal- or lower-scoring principal ("No Better vs. No Change"); average standardized math and reading score. Panel (b): event-study estimate of the demotion spillover for schools whose own principal was *not* summoned (non-reform sample) — receipt of a demoted ex-principal (essentially all outsiders) as a teacher in 2015; average standardized math and reading score. The omitted year is 2014. Standard errors clustered at the school level. Underlying coefficients in Appendix Tables A14 and A12.

Appendices

A Appendix

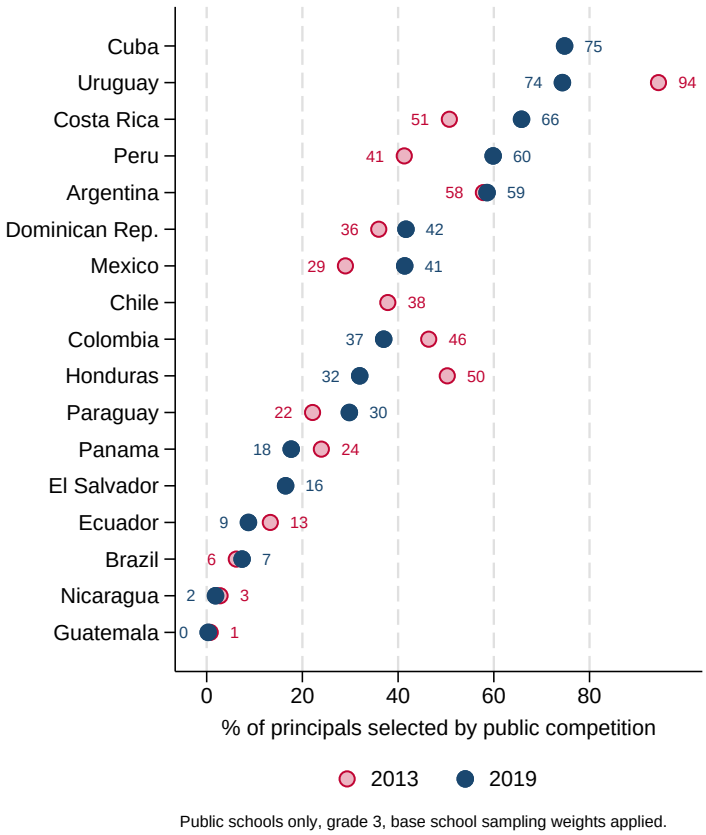
A.1 Additional tables and figures

Table A1: Summary statistics

	Analysis Sample			Spillover Sample			Universe		
	Mean	SD	N	Mean	SD	N	Mean	SD	N
School Characteristics									
Number of Students	239.43	243.60	2,543	213.62	232.32	8,827	114.79	187.33	51,356
Number of Teachers	10.18	8.53	2,534	9.10	8.22	8,803	6.24	8.65	51,356
Share Female	0.49	0.10	2,534	0.49	0.09	8,803	0.48	0.13	51,356
Urban	0.51	0.50	2,543	0.42	0.49	8,827	0.41	0.49	51,356
Resource Index	-0.03	0.83	2,368	-0.01	0.82	8,012	-0.02	0.94	32,999
Share Met Basic Needs (%)	54.34	21.66	2,543	49.71	22.02	8,827	50.05	24.00	51,328
Amenity Index	0.29	0.84	2,543	0.15	0.85	8,827	-0.00	1.01	51,356
Principal Characteristics									
Female	0.33	0.47	2,394	0.36	0.47	4,929	0.33	0.46	15,823
Age	50.28	6.55	2,242	50.02	7.44	4,779	49.27	7.99	15,093
Salary (PEN)	2,723.58	713.72	2,003	2,530.47	773.50	4,216	2,401.64	801.58	14,346
Principal Turnover, 2014–2015	0.60	0.49	2,166	0.73	0.44	4,664	0.71	0.46	12,286
Principal Transfer In, 2014–2015	0.50	0.50	2,084	0.65	0.48	5,140	0.61	0.49	13,475
Principal Exams									
Excepcional: Summoned	1.00	0.00	2,543	0.45	0.50	8,827	0.19	0.39	51,356
Excepcional: Taker	1.00	0.00	2,543	0.29	0.45	8,827	0.12	0.32	51,356
Excepcional: Passed	0.35	0.48	2,543	0.10	0.30	8,827	0.04	0.20	51,356
Acceso: Taker	0.52	0.50	2,543	0.20	0.40	8,827	0.08	0.27	51,356
Acceso: Approved	0.43	0.49	1,329	0.41	0.49	1,784	0.40	0.49	3,962
Teacher Characteristics									
Share Female	0.63	0.27	2,505	0.62	0.27	8,657	0.50	0.36	33,635
Average Age	46.69	5.73	2,448	45.72	5.99	8,484	43.65	7.11	32,863
Average Salary (PEN)	1,669.60	199.08	2,044	1,655.83	230.63	7,372	1,642.70	423.50	30,631
Student Performance									
Avg Raw Math Score (2nd)	553.43	88.79	2,543	543.46	91.10	8,827	525.64	94.84	22,674
Avg Raw Reading Score (2nd)	552.31	58.45	2,543	546.21	61.84	8,827	543.66	67.62	22,674
Promotion Rate	0.90	0.08	2,543	0.90	0.08	8,827	0.86	0.15	51,356
Dropout Rate	0.02	0.03	2,543	0.02	0.03	8,827	0.02	0.05	51,356

Note: Summary statistics measured in 2014 (the year of the reform) at the school level. "Analysis Sample" includes schools whose incumbent principal took the *Excepcional* exam. "Spillover Sample" includes schools used in the demotion spillover analysis. "Universe" includes all primary schools in Peru. Raw student scores are reported in the original ECE scale; standardized scores have been transformed to mean zero, standard deviation one.

Figure A1: How principals get their jobs across Latin America



Note: Data from ERCE. Share of principals selected through different mechanisms across Latin American countries, for the set of countries appearing in both survey waves of ERCE. See Appendix Figure A5 for results by survey year (2013 and 2019).

A.2 The examinations and test validity

The 2014 reform involved two examinations. The *Excepcional* exam, administered in August 2014, was a one-off examination designed to give incumbent principals a chance to retain their positions. The *Acceso* exam, administered in November 2014, was the regular entry examination for principal positions, open to all eligible teachers.

Both examinations assessed candidates on several dimensions. Approximately half of the total exam weight was allocated to instructional and operational planning—the ability to design curricula, organize school activities, and plan pedagogical strategies. The remaining weight was distributed across modules on conflict resolution and human resource management, financial management and resource allocation, and external engagement with the community and education authorities.

It is not clear ex-ante that civil service exams—and especially managerial exams—can accurately measure the multi-faceted skills required for effective school leadership. Good management is a complex construct, only partially codifiable. However, a growing literature suggests that civil service exam performance does predict later job performance in some contexts. Dahis et al. (2025) show that admission exam scores predict later performance among Brazilian bureaucrats. Bertrand et al. (2020) find that entry scores into the Indian Administrative Service predict career advancement. In a setting more directly relevant to ours, Aman-Rana (2025) examines Pakistan's civil service exam and finds limited correlation between exam rank and on-the-job delivery performance—highlighting that this relationship is context-dependent.

In the Peruvian setting, we can estimate the correlation between a principal's exam scores and their school's performance. Table A2 reports correlations between standardized exam scores and several measures of school outcomes, including average student test scores, manager performance evaluations, and student value-added. Panel A shows that principals who scored higher on the reading comprehension component of the *Excepcional* exam led schools with significantly higher student test scores in both math (0.042 SD per exam SD, $p < 0.01$) and reading (0.037 SD, $p < 0.01$). Panels C–E show that *Acceso* exam scores in reading comprehension, management concepts, and education level are all positively and significantly correlated with both the 2018 manager performance evaluation and student value-added in math and reading. Panel F shows that the experience component does not predict outcomes, consistent with the idea that the knowledge-based components of the exam are the most informative.

These results suggest that the Peruvian examinations do capture skills that are linked to school performance, lending credibility to the reform's design—even as the reform's implementation generated the countervailing effects documented in the main text.

A note on combining *Excepcional* and *Acceso* scores. In the broader-turnover analysis (Section 4.4), we classify principal transitions as Better or No Better by comparing exam scores of the incoming and outgoing principal, using the *Excepcional* score when available and falling back

Table A2: Exam score correlations with student outcomes and manager performance

	(1)	(2)	(3)	(4)	(5)
	Student scores '11-'13		Manager Performance	Student value-added '14-'16, '16-'18	
	Mathematics	Reading	Evaluation '18	Mathematics	Reading
A. Extraordinary Contest, Reading Comprehension					
Z(Exam Score)	0.042***	0.037***	0.064**	-0.007	-0.009
	(0.010)	(0.009)	(0.031)	(0.014)	(0.011)
Adjusted R2	0.167	0.214	0.155	0.086	0.053
N Managers	3,729	3,729	2,039	1,293	1,293
Observations	339415	339654	2039	83105	83135
B. Extraordinary Contest, Management Concepts					
Z(Exam Score)	0.047***	0.042***	0.013	0.002	-0.000
	(0.014)	(0.011)	(0.032)	(0.014)	(0.010)
Adjusted R2	0.153	0.186	0.162	0.082	0.050
N Managers	1,913	1,913	1,698	1,070	1,070
Observations	209211	209396	1698	74428	74452
C. Managerial Access Contest, Reading Comprehension					
Z(Exam Score)			0.059**	0.023**	0.018**
			(0.023)	(0.011)	(0.009)
Adjusted R2			0.162	0.067	0.036
N Managers			6,873	3,968	3,968
Observations			6873	245084	245172
D. Managerial Access Contest, Management Concepts					
Z(Exam Score)			0.136***	0.025**	0.025***
			(0.019)	(0.011)	(0.009)
Adjusted R2			0.167	0.067	0.036
N Managers			6,873	3,968	3,968
Observations			6873	245084	245172
E. Managerial Access Contest, Education Level					
Z(Exam Score)			0.095***	0.018***	0.014***
			(0.013)	(0.006)	(0.005)
Adjusted R2			0.168	0.068	0.036
N Managers			6,863	3,961	3,961
Observations			6863	244029	244117
F. Managerial Access Contest, Experience					
Z(Exam Score)			0.006	0.002	0.002
			(0.011)	(0.005)	(0.004)
Observations			6863	244029	244117
Adjusted R2			0.161	0.067	0.036
N Managers			6,863	3,961	3,961
Local Authority FE	✓	✓	✓	✓	✓
School controls	✓	✓	✓	✓	✓
Student controls	✓	✓	✓	✓	✓
Sample	Takers Primary	Takers Primary	Approved All	Approved Primary	Approved Primary

on the *Acceso* score otherwise. The two exams share most of their content blocks (operational planning, conflict resolution, financial management, community engagement) and are scored on similar scales, but they differ in test-taker population (*Excepcional* was a one-off retention exam for incumbents in 2014, whereas *Acceso* is the regular entry competition open to all eligible candidates). The Better/No Better classification is therefore approximate; we treat it as descriptive rather than as a rigorous principal-quality measure.

Construction of the principal-quality score. We build our measure of principal quality from the official records of the 2014 *Excepcional* and *Acceso* examinations administered by MINEDU. Both exams share a common Text Comprehension subtest (*pje_PCT* in the *Acceso* data; PCT in the *Excepcional* data), scored from 0 to 100 in increments of four points with a passing threshold of 60. Candidates who clear that threshold proceed to a Case Solution subtest (*pje_PSC*), also scored from 0 to 100. The *Acceso* version is a single 40-question test scored at 2.5 points per correct answer; the *Excepcional* version is split into a Conflict Resolution case (PSC_PC) and a Pedagogical case (PSC_PP), each worth 50 points, which sum to a total of 100.

For *Acceso* candidates who classify and submit the required documentation, MINEDU also records a Trajectory score (*pje_tot_trayectoria*, 0 to 76) that combines Academic Background, Merits, and Professional Experience. The MINEDU partial score weights the three components,

$$\text{Partial}_{Acceso} = 0.3 \cdot \text{PCT} + 0.5 \cdot \text{PSC} + 0.2 \cdot \text{Trajectory},$$

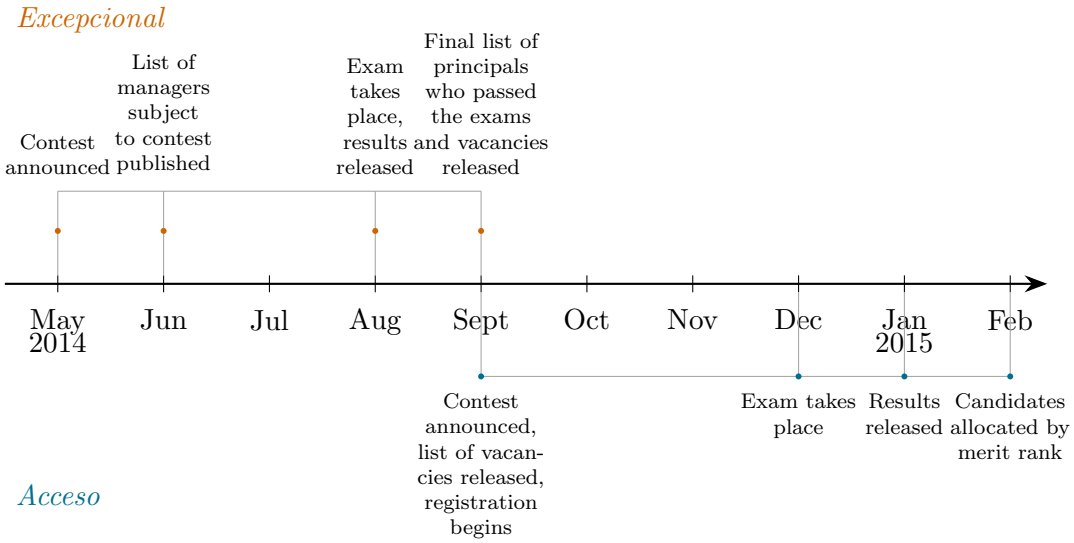
and ranges from 28.15 to 84.35 in our data; the final score adds bonuses for registered disability or armed-forces service and ranges from 28.15 to 92.65. Approval (*SITU_FINAL*) and allocation (*ADJUDICADO*) decisions do not follow a strict score rule: candidates with low final scores are sometimes Approved and Allocated, and candidates with high subtest scores are sometimes neither. This is one reason we do not pursue a regression-discontinuity design around any of the cutoffs. The *Excepcional* exam does not include a Trajectory component, and so has no analogue of the partial or final score; *Excepcional* approval is based on Case Solution alone, with eleven exceptions in our data (four with injunction orders—*medida cautelar*—and seven who failed the *Educación Intercultural Bilingüe* component).

In the heterogeneity analyses we use the Text Comprehension subtest score as our measure of principal quality. Text Comprehension is the only component observed for every candidate who sat either exam, and is therefore the only measure on which we can directly compare incoming and outgoing principals at the same school. The Case Solution score is missing for *Excepcional* candidates who failed the Text Comprehension threshold, and the MINEDU partial and final scores are missing for the large share of *Acceso* candidates without a Trajectory score. When a principal has Text Comprehension scores from both exams—typically incumbents who sat *Excepcional* and

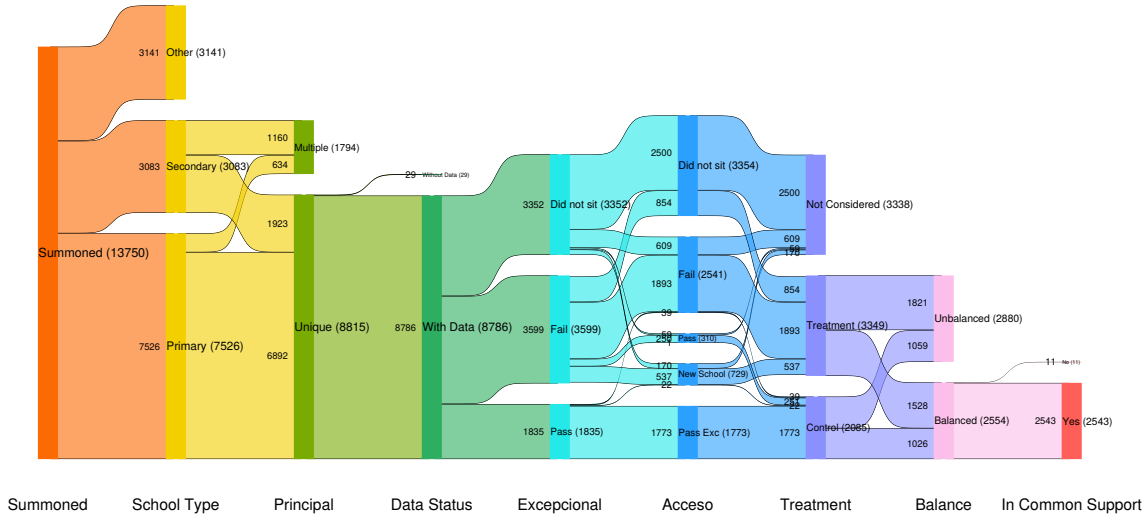
then took *Acceso* after failing—we use the *Excepcional* score; otherwise we use the *Acceso* score.

Figure A2: The 2014 reform: timeline and principal flows

(a) Timeline of the 2014 reform examinations

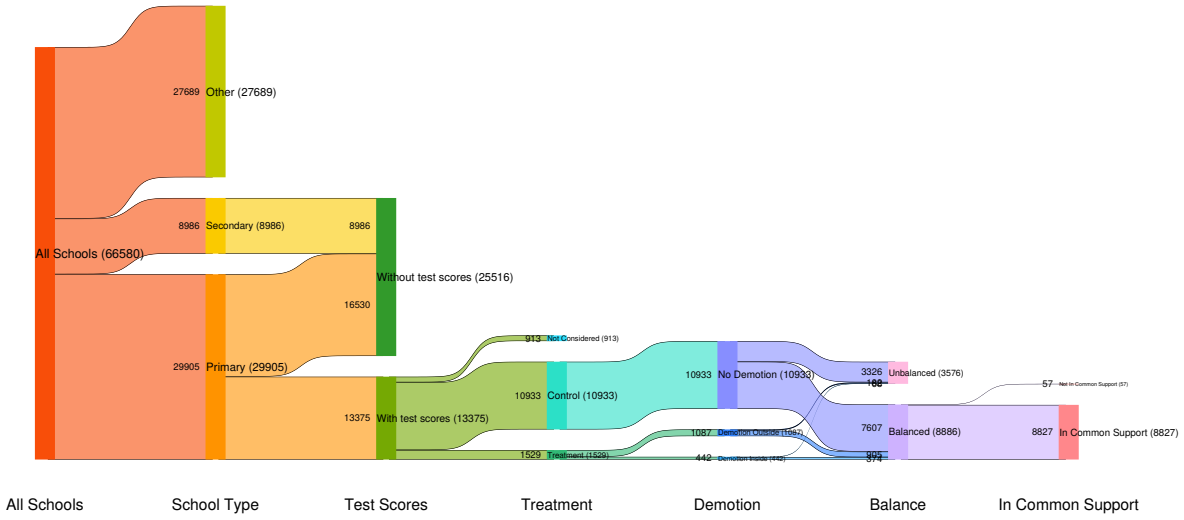


(b) Principal flows through the reform examinations



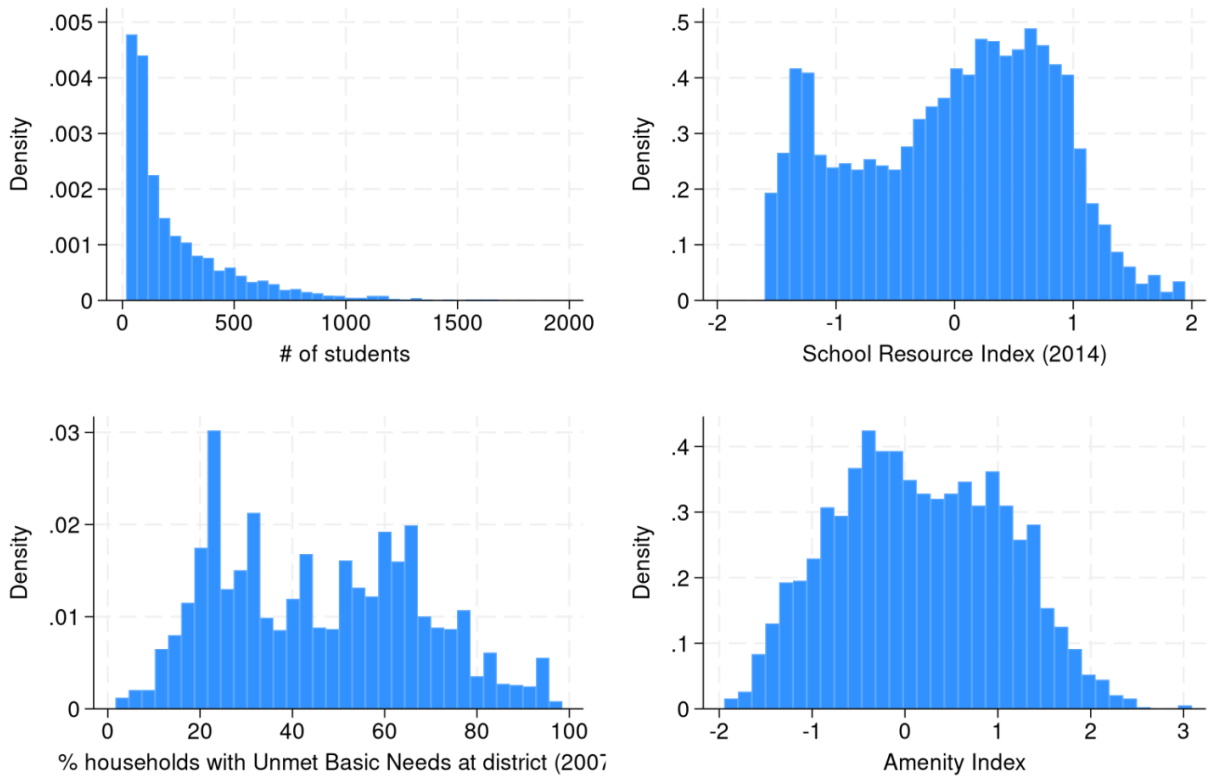
Note: Panel (a): Timeline of the two 2014 examinations. The *Excepcional* process (top track) ran from May to September 2014, giving incumbent principals a chance to retain their positions. The *Acceso* process (bottom track) ran from September 2014 to February 2015 as the regular entry competition open to all eligible candidates. Each track moves through a Classification stage (Reading Comprehension exam, must score 60+ to advance, centralized), and Approval/Ranking stage (Management Cases exam, centralized administration with decentralized scoring) when exams take place. A Job Offer stage and a Job Acceptance stage takes place when the final list of exam-passing principals (*Excepcional*) and candidates order by merit rank *Acceso* are released . Panel (b): Flow of incumbent principals through the 2014 examination process. All 13,750 principals in covered schools were summoned to the *Excepcional* exam. Those who failed or did not take it could attempt the *Acceso* exam. Source: Administrative exam records from MINEDU.

Figure A3: Flows of demoted principals



Note: This figure traces the flow of principals who failed the examinations into teaching positions. Source: Nexus administrative data matched to exam records, MINEDU.

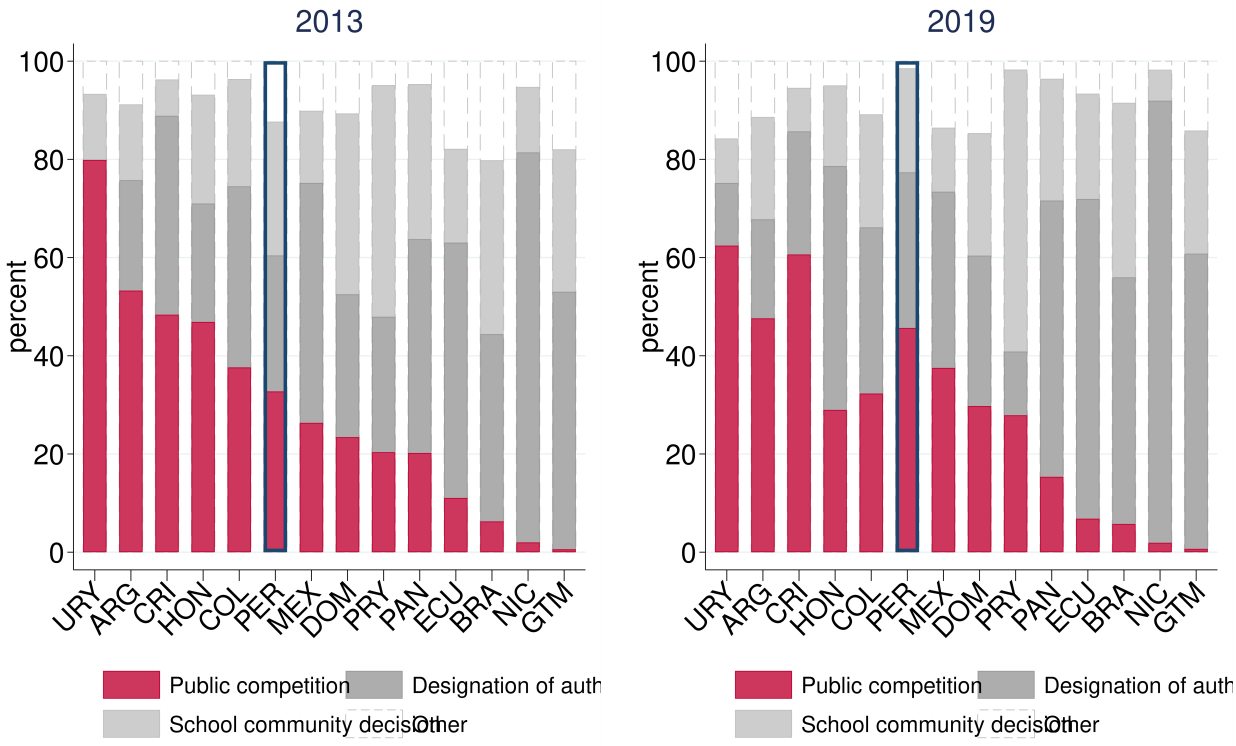
Figure A4: Distribution of amenity index components



Note: Histograms of three of the four inputs to the amenity index used to characterize school context (the urban-vs-rural binary indicator is not plotted) and of the resulting composite. Top-left: school enrollment (number of students). Top-right: school resource infrastructure index, a composite of facility and equipment measures from MINEDU's 2014 school census, standardized to mean zero and unit variance. Bottom-left: share of households at the district level with unmet basic needs (NBI) from the 2007 Population and Housing Census; the amenity index is constructed using its complement, the share with basic needs met. Bottom-right: the resulting amenity index, computed as a standardized weighted average of the four components (urban indicator, log enrollment, the school resource index, and the share of households with basic needs met). Sample: schools in the analysis panel observed in 2014. Sources: MINEDU Nexus and the 2014 school census; INEI 2007 Population and Housing Census.

Figure A5: Principal selection mechanisms by ERCE survey year

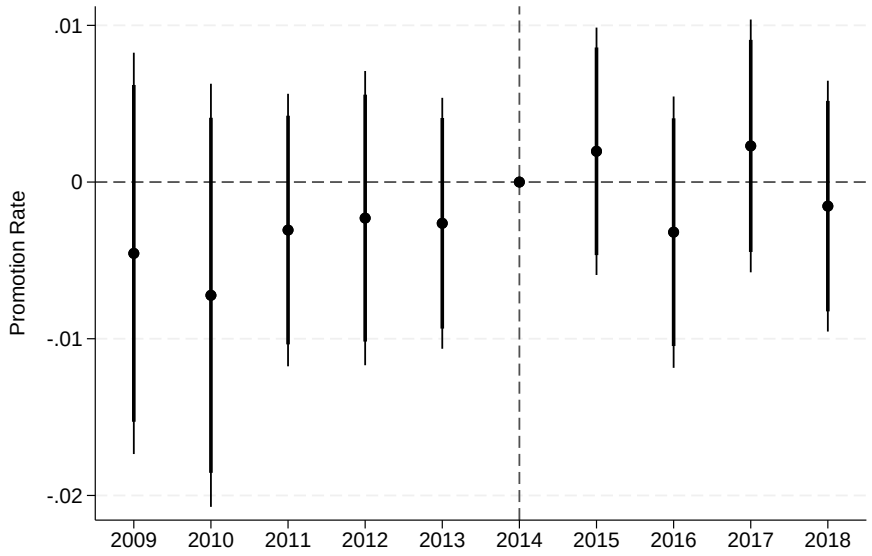
(a) 2013 (b) 2019



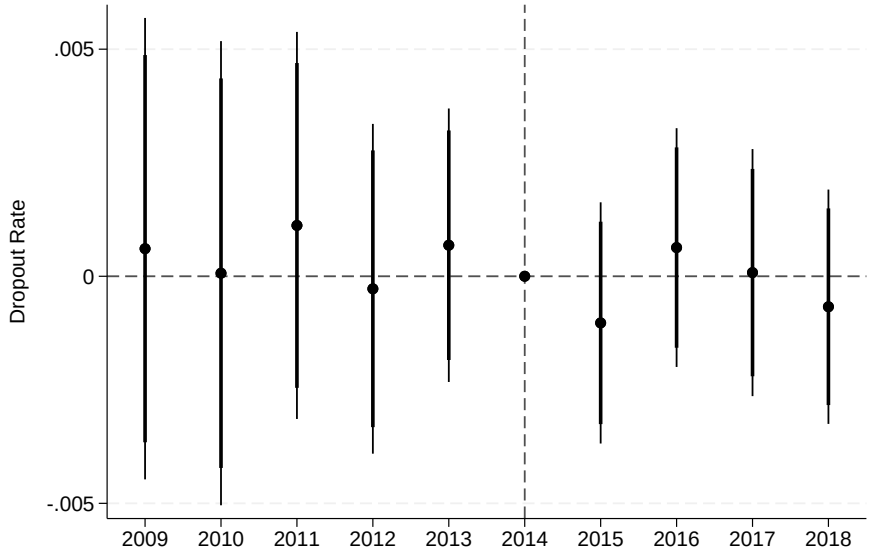
Note: Source: ERCE data. Each panel shows the share of principals selected through different mechanisms, for the set of countries appearing in both survey waves.

Figure A6: Event study: promotion and dropout rates

(a) Promotion rate



(b) Dropout rate



Note: This figure shows event-study estimates of the reform's effect on promotion and dropout rates. See Figure 4 notes for specification details.

Table A3: Heterogeneous effects by school characteristics: math and reading separately

		Amenities Index		School Location		School Size		School Resources		Met Needs		First-Time Princ.	
	(1) Overall	(2) Low	(3) High	(4) Rural	(5) Urban	(6) Small	(7) Large	(8) Low	(9) High	(10) Low	(11) High	(12) Yes	(13) No
Student Scores, Math													
ITT × Post	-0.065**	-0.100**	-0.009	-0.106**	-0.009	-0.121***	-0.018	-0.096**	-0.055	-0.111**	-0.017	-0.069**	-0.071*
	(0.026)	(0.047)	(0.032)	(0.045)	(0.028)	(0.042)	(0.028)	(0.043)	(0.036)	(0.044)	(0.033)	(0.028)	(0.043)
Schools	2,543	1,194	1,194	1,246	1,291	1,276	1,250	1,220	1,237	1,266	1,267	2,143	1,222
School-Year Obs	22,887	10,746	10,746	11,214	11,619	11,484	11,250	10,980	11,133	11,394	11,403	19,287	10,998
R-squared	0.137	0.106	0.217	0.104	0.220	0.105	0.229	0.138	0.143	0.117	0.200	0.136	0.131
Mean Dep. Var (2016)	0.238	-0.075	0.548	-0.068	0.534	-0.035	0.520	0.228	0.256	-0.059	0.538	0.254	0.254
Student Scores, Reading													
ITT × Post	-0.064**	-0.101**	-0.001	-0.107***	-0.011	-0.119***	-0.006	-0.092**	-0.053	-0.108***	-0.024	-0.067**	-0.085**
	(0.025)	(0.043)	(0.031)	(0.041)	(0.027)	(0.040)	(0.025)	(0.038)	(0.035)	(0.038)	(0.035)	(0.026)	(0.041)
Schools	2,543	1,194	1,194	1,246	1,291	1,276	1,250	1,220	1,237	1,266	1,267	2,143	1,222
School-Year Obs	22,887	10,746	10,746	11,214	11,619	11,484	11,250	10,980	11,133	11,394	11,403	19,287	10,998
R-squared	0.153	0.150	0.211	0.150	0.222	0.144	0.227	0.156	0.158	0.164	0.194	0.152	0.152
Mean Dep. Var (2016)	0.086	-0.295	0.466	-0.303	0.461	-0.256	0.438	0.067	0.106	-0.268	0.441	0.105	0.097

Note: IPW-DiD estimates of the average post-reform effect on student achievement by school characteristics, reported separately for math (Panel A) and reading (Panel B). Outcome is the combined 2nd- and 4th-grade standardized score as described in Section 2.3. See Table A11 for specification details. Robust standard errors clustered at the school level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Heterogeneous effects: math and reading separately (2nd grade only, 2009–2016)

		Amenities Index		School Location		School Size		School Resources		Met Needs		First-Time Princ.	
	(1) Overall	(2) Low	(3) High	(4) Rural	(5) Urban	(6) Small	(7) Large	(8) Low	(9) High	(10) Low	(11) High	(12) Yes	(13) No
Student Scores, Math													
ITT × Post	-0.084***	-0.122**	-0.006	-0.132***	-0.012	-0.139***	-0.034	-0.090*	-0.086**	-0.129***	-0.027	-0.090***	-0.096**
	(0.030)	(0.051)	(0.033)	(0.048)	(0.032)	(0.047)	(0.034)	(0.050)	(0.044)	(0.049)	(0.035)	(0.031)	(0.047)
Schools	2,603	1,225	1,227	1,297	1,293	1,291	1,286	1,244	1,267	1,299	1,287	2,194	1,254
School-Year Obs	20,824	9,800	9,816	10,376	10,344	10,328	10,288	9,952	10,136	10,392	10,296	17,552	10,032
R-squared	0.140	0.118	0.199	0.116	0.206	0.108	0.222	0.142	0.144	0.129	0.181	0.137	0.138
Mean Dep. Var (2016)	0.357	0.062	0.658	0.069	0.651	0.063	0.656	0.352	0.367	0.095	0.621	0.367	0.370
Student Scores, Reading													
ITT × Post	-0.085***	-0.124**	-0.015	-0.131***	-0.028	-0.153***	-0.022	-0.087*	-0.093**	-0.145***	-0.031	-0.083***	-0.126***
	(0.029)	(0.050)	(0.031)	(0.048)	(0.030)	(0.047)	(0.030)	(0.047)	(0.042)	(0.048)	(0.037)	(0.031)	(0.047)
Schools	2,603	1,225	1,227	1,297	1,293	1,291	1,286	1,244	1,267	1,299	1,287	2,194	1,254
School-Year Obs	20,824	9,800	9,816	10,376	10,344	10,328	10,288	9,952	10,136	10,392	10,296	17,552	10,032
R-squared	0.155	0.154	0.208	0.156	0.221	0.144	0.228	0.157	0.163	0.168	0.191	0.154	0.152
Mean Dep. Var (2016)	0.148	-0.231	0.534	-0.236	0.537	-0.228	0.528	0.134	0.166	-0.193	0.490	0.163	0.162

Note: This table reports IPW-DiD estimates separately for math (Panel A) and reading (Panel B), restricted to 2nd-grade scores over 2009–2016. See Table A11 notes for specification details.

Table A5: Annual ITT coefficients on student outcomes — overall sample

	Math	Reading	Average	Promotion	Dropout
Panel A: Pre-reform					
ITT × 2009	-0.008	-0.003	-0.006	0.001	-0.005
	(0.052)	(0.050)	(0.049)	(0.003)	(0.007)
ITT × 2010	-0.018	-0.015	-0.017	0.000	-0.007
	(0.053)	(0.048)	(0.048)	(0.003)	(0.007)
ITT × 2011	-0.014	-0.007	-0.011	0.001	-0.003
	(0.049)	(0.047)	(0.046)	(0.002)	(0.004)
ITT × 2012	-0.017	-0.016	-0.017	-0.000	-0.002
	(0.045)	(0.042)	(0.041)	(0.002)	(0.005)
ITT × 2013	-0.004	-0.003	-0.003	0.001	-0.003
	(0.044)	(0.041)	(0.041)	(0.002)	(0.004)
Panel B: Post-reform					
ITT × 2015	-0.118***	-0.115***	-0.116***	-0.001	0.002
	(0.045)	(0.043)	(0.042)	(0.001)	(0.004)
ITT × 2016	-0.080*	-0.078*	-0.079*	0.001	-0.003
	(0.044)	(0.040)	(0.041)	(0.001)	(0.004)
ITT × 2017				0.000	0.002
				(0.001)	(0.004)
ITT × 2018	-0.030	-0.022	-0.026	-0.001	-0.002
	(0.049)	(0.044)	(0.045)	(0.001)	(0.004)
Schools	2,543	2,543	2,543	2,543	2,543
School-Year Obs	22,887	22,887	22,887	25,419	25,419
R-squared	0.567	0.657	0.630	0.456	0.396
Mean Dep. Var (2016)	0.238	0.086	0.162	0.010	0.912

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: Annual ITT×Year coefficients from the event-study specification of equation 3 estimated on the overall ITT sample, reported for each year in the analysis window. Standard errors clustered at the school level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.3 Detailed channel analysis: principal, school, and teacher characteristics

The tables below report the underlying IPW-DiD coefficients summarized in main-text Table 2. Each subgroup row shows the within-subgroup ITT×Post coefficient on the listed outcome, with the "p-value coeff. equality" row reporting the formal test of equality between the contrast pair

Table A6: Annual ITT coefficients on student outcomes — low-amenity schools

	Test Scores (std. dev.)			Rates	
	Math	Reading	Average	Promotion	Dropout
Panel A: Pre-reform					
ITT × 2009	-0.013	-0.014	-0.014	-0.004	-0.000
	(0.088)	(0.089)	(0.086)	(0.010)	(0.004)
ITT × 2010	-0.021	-0.018	-0.020	-0.009	-0.000
	(0.086)	(0.080)	(0.080)	(0.012)	(0.004)
ITT × 2011	-0.019	-0.014	-0.017	-0.005	0.000
	(0.076)	(0.075)	(0.072)	(0.008)	(0.004)
ITT × 2012	-0.023	-0.021	-0.022	-0.003	-0.001
	(0.076)	(0.076)	(0.073)	(0.008)	(0.003)
ITT × 2013	0.005	0.012	0.008	-0.004	0.000
	(0.080)	(0.075)	(0.074)	(0.007)	(0.003)
Panel B: Post-reform					
ITT × 2015	-0.159**	-0.144**	-0.152**	0.001	0.001
	(0.070)	(0.073)	(0.068)	(0.007)	(0.002)
ITT × 2016	-0.122	-0.123*	-0.122*	-0.005	0.003
	(0.076)	(0.073)	(0.072)	(0.008)	(0.002)
ITT × 2017				0.001	0.003
				(0.007)	(0.003)
ITT × 2018	-0.054	-0.063	-0.059	-0.004	0.001
	(0.080)	(0.074)	(0.073)	(0.007)	(0.002)
Schools	1,194	1,194	1,194	1,194	1,194
School-Year Obs	10,746	10,746	10,746	11,932	11,932
R-squared	0.484	0.551	0.533	0.406	0.464
Mean Dep. Var (2016)	-0.075	-0.295	-0.185	0.919	0.010

Note: Annual ITT×Year coefficients on the low-amenity subsample (below median amenity index). See main-text Table 1 for the corresponding DiD summary. Standard errors clustered at the school level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Annual ITT coefficients on student outcomes — high-amenity schools

	Test Scores (std. dev.)			Rates	
	Math	Reading	Average	Promotion	Dropout
Panel A: Pre-reform					
ITT × 2009	0.031	0.022	0.026	-0.001	-0.002
	(0.060)	(0.051)	(0.053)	(0.008)	(0.003)
ITT × 2010	0.025	0.029	0.027	-0.002	-0.001
	(0.063)	(0.054)	(0.056)	(0.008)	(0.003)
ITT × 2011	0.022	0.032	0.027	-0.002	0.000
	(0.064)	(0.055)	(0.058)	(0.005)	(0.002)
ITT × 2012	0.021	0.019	0.020	0.000	-0.001
	(0.053)	(0.042)	(0.045)	(0.006)	(0.002)
ITT × 2013	-0.008	0.002	-0.003	0.000	0.001
	(0.054)	(0.041)	(0.044)	(0.005)	(0.002)
Panel B: Post-reform					
ITT × 2015	0.028	0.012	0.020	0.000	-0.001
	(0.054)	(0.047)	(0.048)	(0.005)	(0.001)
ITT × 2016	-0.019	-0.016	-0.018	0.004	-0.001
	(0.053)	(0.042)	(0.046)	(0.005)	(0.001)
ITT × 2017				0.005	-0.002
				(0.005)	(0.001)
ITT × 2018	0.009	0.054	0.031	0.002	-0.002
	(0.070)	(0.059)	(0.063)	(0.005)	(0.002)
Schools	1,194	1,194	1,194	1,194	1,194
School-Year Obs	10,746	10,746	10,746	11,939	11,939
R-squared	0.648	0.727	0.702	0.424	0.477
Mean Dep. Var (2016)	0.548	0.466	0.507	0.905	0.010

Note: Annual ITT×Year coefficients on the high-amenity subsample (above median amenity index). See main-text Table 1 for the corresponding DiD summary. Standard errors clustered at the school level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

(e.g., Rural vs. Urban).

Table A8: Channel analysis: principal, school and teacher characteristics — overall and by school context

	Principal					School				
	(1) Age	(2) Log Sal	(3) Fem	(4) Test Score	(5) Moved	(6) Demotion	(7) % New Tch	(8) Teacher age	(9) Teacher salary	(10) Teacher fem
ITT (all schools)										
ITT × Post	-4.312***	0.042***	0.037*	17.778***	0.702***	0.080***	-0.001	0.295	0.005	-0.007
	(0.294)	(0.014)	(0.019)	(0.507)	(0.014)	(0.004)	(0.001)	(0.210)	(0.007)	(0.009)
Outcome Mean (2014)	49.662	7.914	0.305	61.872	0.000	0.000	0.001	46.429	7.415	0.620
Schools	1,959	1,666	2,030	2,064	2,079	2,543	2,543	2,543	2,543	2,542
Heterogeneity by Localization										
Rural	-3.440***	-0.039*	-0.018	14.635***	0.726***	0.084***	-0.001	0.227	0.011	-0.005
	(0.409)	(0.023)	(0.025)	(0.725)	(0.019)	(0.007)	(0.002)	(0.348)	(0.012)	(0.015)
Schools	894	717	989	934	1,060	1,246	1,246	1,246	1,246	1,245
Urban	-5.235***	0.116***	0.101***	20.589***	0.665***	0.068***	-0.000	0.366	0.003	-0.013
	(0.421)	(0.018)	(0.030)	(0.703)	(0.023)	(0.006)	(0.001)	(0.226)	(0.006)	(0.010)
Schools	1,060	945	1,036	1,124	1,014	1,291	1,291	1,291	1,291	1,291
p-value coeff. equality	0.002	0.000	0.002	0.000	0.042	0.088	0.883	0.737	0.542	0.676
Heterogeneity by Amenity										
Low Amenity	-3.312***	-0.018	-0.013	14.983***	0.721***	0.083***	-0.001	0.134	0.000	-0.007
	(0.404)	(0.024)	(0.026)	(0.722)	(0.020)	(0.008)	(0.002)	(0.345)	(0.011)	(0.015)
Schools	883	719	941	906	1,010	1,194	1,194	1,194	1,194	1,193
High Amenity	-5.276***	0.100***	0.111***	20.423***	0.648***	0.065***	-0.001	0.220	0.007	-0.003
	(0.450)	(0.021)	(0.030)	(0.753)	(0.025)	(0.006)	(0.001)	(0.216)	(0.006)	(0.007)
Schools	960	840	960	1,027	938	1,194	1,194	1,194	1,194	1,194
p-value coeff. equality	0.001	0.000	0.002	0.000	0.023	0.060	0.965	0.834	0.584	0.796
Heterogeneity by Resources										
Low Resources	-4.003***	0.027	0.018	16.844***	0.716***	0.075***	-0.001	0.182	-0.001	0.002
	(0.412)	(0.018)	(0.027)	(0.722)	(0.020)	(0.007)	(0.001)	(0.317)	(0.009)	(0.016)
Schools	945	807	976	996	993	1,220	1,220	1,220	1,220	1,219
High Resources	-4.319***	0.066***	0.061**	18.811***	0.690***	0.086***	-0.001	0.287	0.009	-0.017
	(0.405)	(0.024)	(0.027)	(0.731)	(0.021)	(0.006)	(0.002)	(0.331)	(0.011)	(0.013)
Schools	948	790	985	996	1,017	1,237	1,237	1,237	1,237	1,237
p-value coeff. equality	0.585	0.192	0.263	0.056	0.380	0.233	0.944	0.818	0.451	0.345
Heterogeneity by School Size										
Small school (17 - 136 students)	-3.606***	-0.014	-0.012	15.938***	0.711***	0.088***	-0.001	0.333	0.009	-0.020
	(0.389)	(0.023)	(0.025)	(0.722)	(0.019)	(0.006)	(0.001)	(0.350)	(0.012)	(0.015)
Schools	927	742	1,003	959	1,067	1,276	1,276	1,276	1,276	1,275

Table A9: Channel analysis: heterogeneity by school resources, basic needs, principal quality, and compliance

	Principal					School				
	(1) Age	(2) Log Sal	(3) Fem	(4) Test Score	(5) Moved	(6) Demotion	(7) % New Tch	(8) Teacher age	(9) Teacher salary	(10) Teacher fem
Heterogeneity by Principal Quality										
No Better Principal	-0.757	-0.128***	0.015	-4.221***	0.423***	0.042***	-0.002	-0.407	0.002	-0.021
	(0.583)	(0.027)	(0.033)	(0.507)	(0.034)	(0.009)	(0.002)	(0.396)	(0.012)	(0.017)
Schools	916	776	976	1,025	1,009	1,025	1,025	1,025	1,025	1,024
Better Principal	-4.578***	0.076***	0.043**	21.048***	0.701***	0.094***	-0.001	0.404*	0.003	-0.006
	(0.326)	(0.014)	(0.021)	(0.481)	(0.016)	(0.005)	(0.001)	(0.223)	(0.007)	(0.009)
Schools	1,631	1,410	1,700	1,802	1,674	1,802	1,802	1,802	1,802	1,801
p-value coeff. equality	0.000	0.000	0.475	0.000	0.000	0.000	0.752	0.074	0.964	0.417
Heterogeneity by Compliance Type										
Defier	1.670***	-0.332***	-0.005*	-0.031	-0.024***	-0.010	-0.003	0.244	-0.000	-0.058**
	(0.640)	(0.028)	(0.003)	(0.095)	(0.007)	(0.009)	(0.002)	(0.596)	(0.019)	(0.029)
Schools	732	613	792	847	822	905	905	905	905	904
Complier	-4.762***	0.076***	0.038*	19.492***	0.783***	0.093***	-0.000	0.239	0.004	-0.002
	(0.307)	(0.015)	(0.021)	(0.537)	(0.014)	(0.004)	(0.001)	(0.210)	(0.007)	(0.009)
Schools	1,846	1,551	1,901	1,905	1,950	2,358	2,358	2,358	2,358	2,357
p-value coeff. equality	0.000	0.000	0.038	0.000	0.000	0.000	0.260	0.994	0.823	0.066

Note: IPW-DiD estimates split by school resources (low/high), district basic-needs-met (above/below 56%), replacement principal quality (Better vs. No Better), and compliance type (Complier vs. Defier). Robust standard errors in parentheses. "p-value coeff. equality" tests whether the contrast pair's coefficients are equal. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A10: Channel analysis: heterogeneity by principal age and experience

	Principal					School				
	(1) Age	(2) Log Sal	(3) Fem	(4) Test Score	(5) Moved	(6) Demotion	(7) % New Tch	(8) Teacher age	(9) Teacher salary	(10) Teacher fem
Heterogeneity by Principal Age										
Older Principal	3.724***	-0.044**	-0.007	12.150***	0.627***	0.085***	-0.004*	0.072	-0.010	-0.012
	(0.317)	(0.020)	(0.029)	(0.858)	(0.025)	(0.007)	(0.002)	(0.347)	(0.011)	(0.015)
Schools	1,159	966	1,195	1,212	1,228	1,261	1,261	1,261	1,261	1,260
Young Principal	-8.123***	0.048***	0.062***	18.240***	0.662***	0.085***	-0.001	0.270	0.009	-0.004
	(0.273)	(0.018)	(0.023)	(0.617)	(0.019)	(0.006)	(0.001)	(0.222)	(0.007)	(0.010)
Schools	1,548	1,276	1,568	1,563	1,529	1,649	1,649	1,649	1,649	1,648
p-value coeff. equality	0.000	0.001	0.061	0.000	0.260	0.993	0.182	0.631	0.145	0.642
Heterogeneity by Principal Experience										
Experienced Principal	-2.426***	-0.037*	0.036	10.962***	0.574***	0.068***	-0.001	0.139	0.010	-0.010
	(0.497)	(0.021)	(0.029)	(0.858)	(0.027)	(0.007)	(0.001)	(0.337)	(0.010)	(0.014)
Schools	1,086	910	1,139	1,206	1,155	1,222	1,222	1,222	1,222	1,221
First Time Principal	-4.697***	0.037**	0.037*	18.815***	0.686***	0.076***	-0.001	0.315	0.005	-0.011
	(0.325)	(0.017)	(0.022)	(0.573)	(0.016)	(0.005)	(0.001)	(0.209)	(0.007)	(0.009)
Schools	1,618	1,373	1,695	1,688	1,764	2,143	2,143	2,143	2,143	2,142
p-value coeff. equality	0.000	0.005	0.983	0.000	0.000	0.322	0.915	0.657	0.712	0.924

Note: IPW-DiD estimates split by principal age (Young vs. Older) and by principal experience (First Time vs. Experienced). Robust standard errors in parentheses. "p-value coeff. equality" tests whether the contrast pair's coefficients are equal. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A.4 Heterogeneity by subgroup-specific reweighted samples

The main-text Table 3 uses a common sample with average IPW weights and adds a subgroup dummy and its interaction with ITT×Post, so the subgroup contrasts are directly comparable. The trade-off is that the IPW weights are not specifically tuned to each subgroup, so the control group for any given subgroup may not be optimally balanced. As a robustness check we re-estimate the heterogeneity analysis the other way: we re-fit the propensity score and IPW weights separately within each subgroup (e.g., low- vs. high-amenity schools) and report the within-subgroup ITT×Post coefficient. This approach maximizes within-subgroup covariate balance but means the resulting coefficients are not strict arithmetic decompositions of the full-sample ITT effect, since each subgroup column is estimated on a different reweighted sample.

Table A11 reports these subgroup-specific reweighted estimates for the same dimensions used in main-text Table 3. The magnitudes are very close to those implied by the common-sample interaction regressions: for example, the Low-Amenity ITT effect is -0.100 here vs. -0.111 implied by the main-text table, and the Rural ITT effect is -0.107 here vs. -0.112 implied by the main-text interaction. These differences (under ~ 0.02 standard deviations) suggest that the cross-subgroup comparability we gain by using a common sample in the main text does not come at a meaningful cost to within-subgroup balance.

Table A11: Heterogeneous effects by school characteristics

		Amenities Index		School Location		School Size		School Resources		Met Needs		Principal Experience		Principal Quality	
	(1) Overall	(2) Low	(3) High	(4) Rural	(5) Urban	(6) Small	(7) Large	(8) Low	(9) High	(10) Low	(11) High	(12) Novice	(13) Experienced	(14) No Better	(15) Better
Student Scores, Avg. Math and Reading															
ITT × Post	-0.065***	-0.100**	-0.005	-0.107**	-0.010	-0.120***	-0.012	-0.094**	-0.054	-0.109***	-0.020	-0.068***	-0.078*	-0.080	-0.049*
	(0.025)	(0.044)	(0.031)	(0.041)	(0.027)	(0.040)	(0.025)	(0.039)	(0.035)	(0.039)	(0.033)	(0.026)	(0.040)	(0.052)	(0.027)
Schools	2,543	1,194	1,194	1,246	1,291	1,276	1,250	1,220	1,237	1,266	1,267	2,143	1,222	1,025	1,802
School-Year Obs	22,887	10,746	10,746	11,214	11,619	11,484	11,250	10,980	11,133	11,394	11,403	19,287	10,998	9,225	16,218
R-squared	0.150	0.133	0.219	0.133	0.226	0.130	0.232	0.153	0.156	0.146	0.203	0.149	0.147	0.135	0.166
Mean Dep. Var (2016)	0.162	-0.185	0.507	-0.185	0.497	-0.146	0.479	0.148	0.181	-0.164	0.490	0.180	0.175	0.223	0.202

Note: IPW-DiD estimates of the average post-reform effect on student achievement (average of math and reading, 2nd grade) by school characteristics. *Variable definitions: Amenities Index* is a standardized weighted index (computed via the `swindex` command in Stata) combining urban location, log 2014 enrollment, the school resource index, and the share of district households with basic needs met. *School Location* is urban vs. rural per MINEDU's classification. *School Size* is small (17–136 students) vs. large (> 136), split at the 2014 enrollment median. *School Resources* is a composite index of school facilities and infrastructure, split at the 2014 median. *Met Needs* is the district-level share of households with basic needs met (1 minus the NBI rate), split at the median (56%). Low/High refers to below/above the median of each characteristic. All specifications include school and year fixed effects. Standard errors clustered at the school level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.5 Demotion and turnover spillovers: additional detail

Table A12: Demotion spillover for non-summoned schools: event study

	Test Scores (std. dev.)			Rates	
	Math	Reading	Average	Promotion	Dropout
Panel A: Pre-reform					
Demotion × 2009	-0.038	-0.023	-0.030	0.001	-0.007
	(0.055)	(0.053)	(0.052)	(0.003)	(0.007)
Demotion × 2010	-0.046	-0.027	-0.037	0.001	-0.006
	(0.058)	(0.053)	(0.053)	(0.003)	(0.009)
Demotion × 2011	-0.036	-0.017	-0.027	0.001	-0.005
	(0.052)	(0.049)	(0.048)	(0.002)	(0.005)
Demotion × 2012	-0.045	-0.024	-0.034	0.000	-0.004
	(0.051)	(0.049)	(0.047)	(0.002)	(0.006)
Demotion × 2013	-0.040	-0.025	-0.033	0.000	-0.004
	(0.049)	(0.044)	(0.044)	(0.002)	(0.004)
Panel B: Post-reform					
Demotion × 2015	-0.115**	-0.119**	-0.117**	0.001	-0.001
	(0.053)	(0.047)	(0.048)	(0.001)	(0.004)
Demotion × 2016	-0.051	-0.061	-0.056	-0.000	-0.001
	(0.049)	(0.042)	(0.044)	(0.001)	(0.005)
Demotion × 2017				0.002	-0.008*
				(0.001)	(0.005)
Demotion × 2018	-0.059	-0.027	-0.043	0.001	-0.001
	(0.056)	(0.049)	(0.050)	(0.001)	(0.005)
Schools	4,677	4,677	4,677	4,677	4,677
School-Year Obs	42,093	42,093	42,093	46,746	46,746
R-squared	0.555	0.662	0.627	0.465	0.380
Mean Dep. Var (2016)	0.076	-0.087	-0.006	0.011	0.913

Note: Year-by-year event-study estimates underlying the pooled non-summoned demotion effect shown in Figure 7 Panel (b). Sample restricted to schools whose own principal was *not* summoned for the *Excepcional* examination (the complement of the decomposition sample in Table 4). Treatment is receipt of a demoted ex-principal as a teacher in 2015; 437 treated schools. IPW-weighted with school and year fixed effects; the omitted year is 2014 and 2017 test scores were not collected. Standard errors clustered at the school level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A13: Outsider demotion by own principal-change status

	Test scores (SD)			Rates	
	Math	Reading	Avg	Promotion	Dropout
Panel A: Outsider demotion (reform, principal changed)					
Control group: reform schools with no principal change nor newly demoted teachers					
Group × Post	-0.058*	-0.048	-0.053	-0.001	-0.000
	(0.035)	(0.033)	(0.033)	(0.004)	(0.002)
Treatment N	329	329	329	329	329
Control N	1030	1030	1030	1030	1030
Post-reform first year effect	-0.129*	-0.128**	-0.128**	0.000	0.003*
Panel B: Outsider demotion (reform, principal stayed)					
Control group: reform schools with no principal change nor newly demoted teachers					
Group × Post	-0.073	-0.054	-0.064	-0.004	0.000
	(0.052)	(0.048)	(0.048)	(0.006)	(0.002)
Treatment N	150	150	150	150	150
Control N	1033	1033	1033	1033	1033
Post-reform first year effect	-0.082	-0.061	-0.072	-0.003	0.001
School FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes

Reform outsider demotions split by whether the school's own principal also changed in 2015 (Panel A) or stayed (Panel B); together they decompose the combined outsider group in main-text Table 4 (Panel B). Each cell is the Group×Post coefficient from a difference-in-differences against the reform pure control (reform schools with no principal change and no newly demoted teacher), with school and year fixed effects, group-specific inverse-probability weights, and standard errors clustered by school. The italicized *Post-reform first year effect* is the 2015 event-study coefficient (the group×2015 term from `i.treat##ib2014.year`, with 2014 omitted). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A14: Event study on broader turnover

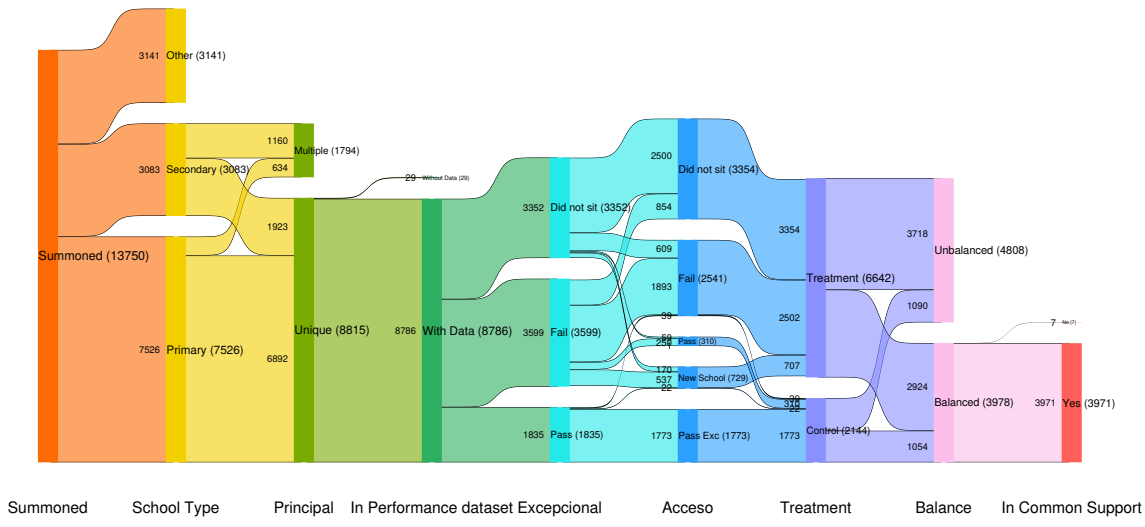
	Math			Reading			Avg Math Read		
	All	No Better	Better	All	No Better	Better	All	No Better	Better
Panel A: Pre-reform									
Turnover × 2009	-0.003	-0.051	0.033	-0.022	-0.022	0.022	-0.012	-0.037	0.027
	(0.043)	(0.064)	(0.045)	(0.040)	(0.059)	(0.042)	(0.040)	(0.059)	(0.042)
Turnover × 2010	0.012	-0.055	0.036	-0.016	-0.027	0.012	-0.002	-0.041	0.024
	(0.041)	(0.072)	(0.044)	(0.037)	(0.064)	(0.040)	(0.037)	(0.065)	(0.040)
Turnover × 2011	0.019	-0.030	0.027	-0.003	0.003	0.017	0.008	-0.013	0.022
	(0.038)	(0.065)	(0.042)	(0.034)	(0.055)	(0.038)	(0.035)	(0.057)	(0.038)
Turnover × 2012	0.022	-0.037	0.039	0.000	-0.007	0.022	0.011	-0.022	0.031
	(0.036)	(0.061)	(0.039)	(0.033)	(0.057)	(0.036)	(0.033)	(0.056)	(0.035)
Turnover × 2013	0.016	-0.041	0.031	-0.006	-0.008	0.013	0.005	-0.024	0.022
	(0.035)	(0.061)	(0.039)	(0.032)	(0.053)	(0.035)	(0.032)	(0.055)	(0.035)
Panel B: Post-reform									
Turnover × 2015	-0.025	-0.092	0.024	-0.047	-0.080	-0.005	-0.036	-0.086	0.010
	(0.038)	(0.066)	(0.041)	(0.036)	(0.060)	(0.038)	(0.035)	(0.061)	(0.038)
Turnover × 2016	0.042	-0.004	0.073*	0.001	0.005	0.039	0.021	0.001	0.056
	(0.035)	(0.059)	(0.038)	(0.032)	(0.049)	(0.033)	(0.032)	(0.052)	(0.034)
Turnover × 2017									
Turnover × 2018	0.089**	0.037	0.140***	0.066*	0.053	0.126***	0.078**	0.045	0.133***
	(0.040)	(0.066)	(0.044)	(0.035)	(0.057)	(0.039)	(0.036)	(0.059)	(0.040)
Schools	4,068	1,497	2,465	4,068	1,497	2,465	4,068	1,497	2,465
School-Year Obs	36,612	13,473	22,185	36,612	13,473	22,185	36,612	13,473	22,185
R-squared	0.574	0.576	0.586	0.673	0.678	0.678	0.643	0.643	0.650
Mean Dep. Var (2016)	0.161	0.232	0.235	0.006	0.080	0.087	0.083	0.156	0.161

Note: Event-study estimates of the relationship between principal turnover and student achievement on the broader turnover sample defined in Section 2.3 (4,165 schools with available student-outcome data, drawn from the 6,898 schools with classifiable 2014–2015 transitions). Each outcome panel (Math, Reading, Avg Math+Reading) reports three columns: All turnover schools pooled, and the Better and No Better partitions defined by comparing incoming and outgoing principals' exam scores (Excepcional if available, Acceso otherwise; see Section A.2 for a note on the validity of mixing the two exam regimes). The coefficient "Turnover×Year" captures the differential trend in achievement

A.6 ITT analysis on the extended sample including non-takers

The main-text ITT analysis restricts the treated group to schools whose incumbent principal sat the *Excepcional* examination. This restriction allows us to observe both incoming and outgoing principal exam scores. Conceptually, however, the reform's "treatment" is the offer to leave or retain the post, which applied to all incumbent principals regardless of whether they ultimately sat the exam; some incumbents declined to test at all. In this appendix we re-estimate the average ITT effect on a extended sample that adds these non-takers to the treated group, while keeping the same set of control schools. Figure A7 reproduces the principal-flow diagram on this extended sample.

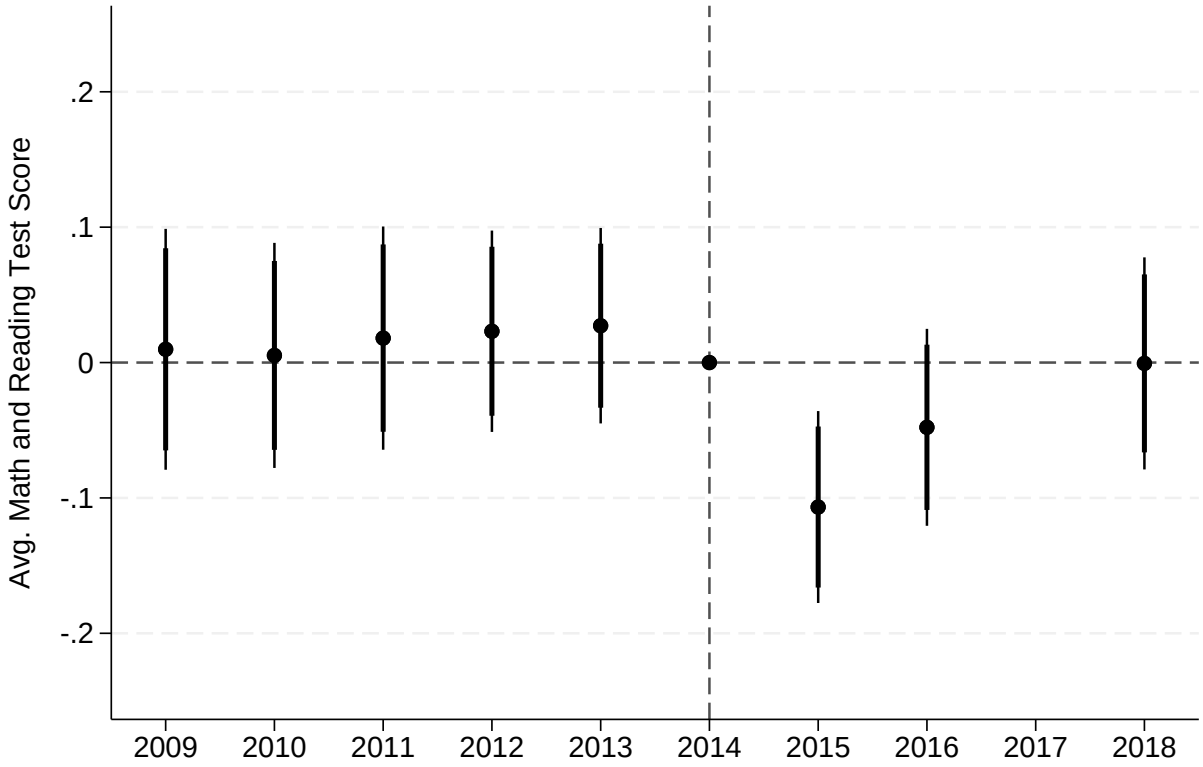
Figure A7: Principal flows through the reform (extended ITT sample)



Note: Sankey diagram of principal transitions in the extended ITT sample, which adds to the main-text treated group those schools whose incumbent principal did not sit the *Excepcional* exam. Control schools are unchanged.

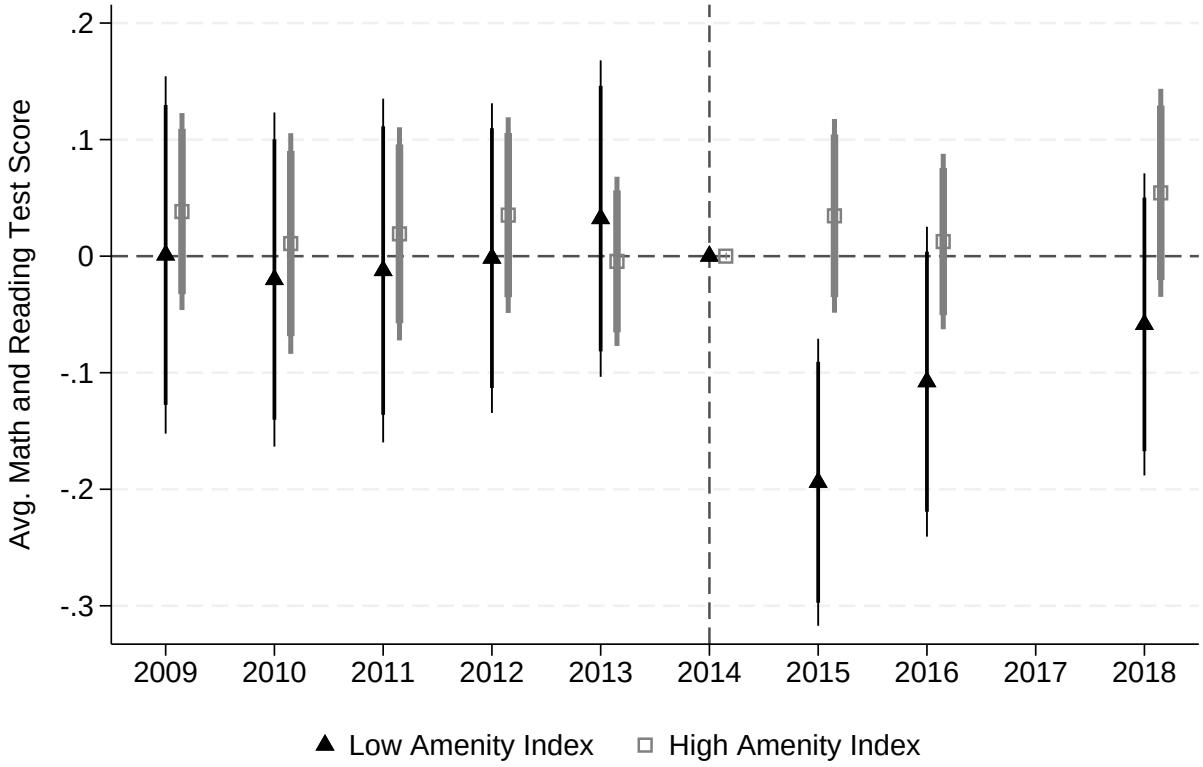
The event-study estimates on this extended sample are shown in Figure A8, with the corresponding leads-and-lags coefficients in Table A15. The pattern is consistent with the main-text analysis: there is a statistically significant decline in student achievement in the first year after the reform (2015), and the effect attenuates in the second year (2016). Under the conceptual framework of Section 3, this evidence is consistent with the non-taker incumbents having been even more negatively selected ex-ante than the failed-takers: their replacements appear to have been "good enough" to overcome the disruption cost from the principal turnover by the second post-reform year.

Figure A8: Event study on the extended ITT sample: average math and reading



Note: Event-study estimates of the reform's effect on average student achievement (combined 2nd- and 4th-grade math and reading) on the extended ITT sample, which adds non-takers to the treated group. Specification, weights, and standard-error treatment match the main-text event study (equation 3). The omitted year is 2014.

Figure A9: Event study on the extended ITT sample: by amenity index



Note: Event-study estimates on the extended ITT sample (including non-takers), stratified by school amenity index (low vs. high, split at the median). Specification, weights, and standard-error treatment match the main-text event study. The omitted year is 2014.

Table A15: ITT effects on student outcomes (extended ITT sample)

	Test Scores (std. dev.)			Rates	
	Math	Reading	Average	Promotion	Dropout
Overall					
ITT × Post	-0.066***	-0.062***	-0.069***	0.002	-0.000
	(0.022)	(0.023)	(0.022)	(0.003)	(0.001)
Schools	3,971	3,971	3,971	3,971	3,971
School-Year Obs	35,739	35,739	35,739	39,697	39,697
R-squared	0.149	0.135	0.153	0.114	0.195
Mean Dep. Var (2016)	0.136	0.210	0.061	0.912	0.011
Low Amenity Index					
ITT × Post	-0.120***	-0.112***	-0.128***	0.000	0.002
	(0.038)	(0.041)	(0.038)	(0.004)	(0.002)
Schools	1,867	1,867	1,867	1,867	1,867
School-Year Obs	16,803	16,803	16,803	18,660	18,660
R-squared	0.131	0.102	0.149	0.139	0.205
Mean Dep. Var (2016)	-0.201	-0.091	-0.312	0.919	0.010
High Amenity Index					
ITT × Post	0.017	0.015	0.019	0.002	0.001
	(0.025)	(0.026)	(0.025)	(0.003)	(0.001)
Schools	1,855	1,855	1,855	1,855	1,855
School-Year Obs	16,695	16,695	16,695	18,546	18,546
R-squared	0.211	0.208	0.206	0.089	0.188
Mean Dep. Var (2016)	0.485	0.522	0.447	0.904	0.011

Note: IPW-weighted difference-in-differences estimates of post-reform ITT effects on student outcomes, on the extended sample that includes non-takers in the treated group. Overall, and split at the median of the school amenity index. Standard errors clustered at the school level in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.