Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from Rwandan primary schools

Clare Leaver[*], Owen Ozier[†], Pieter Serneels[‡], and Andrew Zeitlin[§]

This version: September 13, 2019
PRELIMINARY AND INCOMPLETE

Abstract

Accumulating evidence suggests that performance pay can elicit greater effort from incumbent civil servants, but less is known about how these contracts affect the composition of the public sector workforce. We provide experimental evidence of the impact of performance pay on both the compositional and effort margins. In partnership with the Government of Rwanda, we implemented a 'pay-for-percentile' scheme (Barlevy and Neal 2012) in a novel two-tier experimental design. In the first tier, we randomly assigned teacher labor markets to either performance pay or fixed-wage contracts identical in expected value. In the second tier, we implemented a 'surprise', school-level re-randomization, allowing us to separately identify the compositional effects of *advertised* performance pay and the effort effects of *experienced* performance pay. Our pre-analysis plan sets out a theoretical framework that helps to define a set of hypotheses, and conducts simulations on blinded data to develop high-powered tests. We find that performance pay did change the composition of the teaching workforce, drawing in individuals who were more money-oriented, as measured by a framed Dictator Game. But these recruits were not less effective teachers—if anything the reverse. On the effort margin, we observe substantial and statistically significant gains in teacher value added, mirrored in positive effects on teacher presence and observed pedagogy in the classroom. In Year 2, we estimate the total effect of performance pay, across compositional and effort margins, to be 0.21 standard deviations of pupil learning. One quarter of this impact can be attributed to selection at the recruitment stage, with the remaining three-quarters arising from increased effort.

[*]Blavatnik School of Government, University of Oxford, CEPR, and RISE
[†]World Bank Development Research Group, BREAD, and IZA
[‡]School of International Development, University of East Anglia, IZA, RISE and EGAP
[§]McCourt School of Public Policy, Georgetown University, CGD, and IGC

1 Introduction

The ability to recruit, elicit effort from, and retain civil servants is a central issue for any government. This is particularly true in the context of a sector such as education where people—that is human rather than physical resources—play a key role. Equipping schools with effective teachers generates private returns for students through learning gains, educational attainment, and higher earnings (Chetty et al., 2014a,b), and also social returns as improved skills in the labour force drive economic performance (Hanushek and Woessmann, 2012).

So how can governments best develop people management practices to ensure that schools are equipped with effective teachers? One policy that has been discussed in this context is "pay-for-performance". In its most extreme form, this involves teachers receiving 'pay' as an unconsolidated bonus when their 'performance', measured on a narrow set of objective dimensions, meets some pre-specified criteria, for instance exceeding an absolute level or a relative target. The measured dimensions of performance typically include teacher value-added (constructed from student learning outcomes), and/or teacher inputs such as presence and conduct in the school and classroom.

This form of pay-for-performance divides opinion. Opponents point to public administration, social psychology and, more recently, behavioral economics (Bénabou and Tirole, 2003; Delfgaauw and Dur, 2007) to argue that it will have negative effects on both compositional and effort margins. In short, such contracts are thought to: recruit the wrong types—individuals who are somehow "in it for the money"; lower effort by reducing intrinsic motivation; and fail to retain the right types—good teachers become de-motivated and quit. By contrast, proponents point to classic economic contract theory (Lazear, 2003; Rothstein, 2015) and evidence from private-sector employees in jobs with readily measurable output Lazear (2000) to argue that this form of pay-for-performance will have positive effects on both margins. Under this view, such contracts: recruit the right types—individuals who anticipate performing well in the classroom; raise effort by increasing extrinsic motivation; and retain the right types—good teachers feel rewarded and stay put.

This paper seeks to inject new evidence into this debate. We provide what is to our knowledge the first prospective, experimental evaluation that examines both the compositional and effort margins of pay-for-performance for civil servants. As described below, we implement a novel, two-tiered experiment that allows us to separately identify responses along these margins.

We worked with the Rwanda Education Board and Ministry of Education to design a pay-for-performance (hereafter P4P) contract based on a *pay-for-percentile* system that makes student performance at all levels relevant to teacher rewards (Barlevy and Neal, 2012). Building on extensive consultations and a pilot year, this P4P contract rewards the top 20 percent of teachers with extra pay, over and above their usual salary, using a metric that equally weights learning outcomes in teachers' classrooms alongside three measures of teachers' inputs into the classroom (presence, lesson planning, and observed pedagogy).

Using this contract, our two-tiered experiment (Ashraf et al., 2010; Cohen and Dupas, 2010; Karlan and Zinman, 2009) first randomly assigns labor markets to either P4P or expected-value-equivalent fixed-wage (hereafter FW) *advertisements*, and then uses a surprise re-randomization of *experienced* contracts at the school level to enable estimation of pure compositional effects within each realized contract type. To elaborate, in the first stage—undertaken during recruitment for teacher placements in the 2016 school year—we randomly assigned labor markets to either P4P or FW contracts. Teacher labor markets are defined at the subject by district level, and we conducted the experiment in six districts (18 labor markets) which, together, cover more than half the upper-primary teacher hiring lines in the 2016 school year. We then recruited into the study all primary schools that received such a teacher to fill an upper-primary teaching post (a total of 164 schools). In the second stage of our experiment—undertaken once 2016 teacher placements had been finalized—we randomly re-assigned each of these 164 study schools in their entirety to either P4P or FW contracts; all teachers, including both newly placed recruits and incumbents, who taught core-curricular classes to upper-primary students were eligible for the relevant contracts. We offered a signing bonus to ensure that no recruit, regardless of her belief about the probability of winning, could be made worse off by the re-randomization and, consistent with this, no one turned down their (re-)randomized contract. Incentives were in place for two years, enabling us to study retention as well as to estimate higher-powered tests of effects using outcomes from both years.

Our main findings are as follows. First, in terms of recruitment, advertised P4P contracts did not change the distribution of measured teacher skill (Teacher Training College final exam score) either in the applicant pool as a whole, or among new hires in particular. This is estimated sufficiently precisely to rule out even small negative effects of P4P recruitment on this skill measure. P4P contracts did, however, select teachers who contributed less in a framed Dictator Game played

at baseline to measure intrinsic motivation. Yet in spite of this, teachers recruited under P4P were at least as effective in promoting learning as were those recruited under FW contracts (holding experienced contracts constant). True, P4P contracts did draw in individuals who were more money-motivated but these recruits were *not* less effective teachers, if anything the reverse.

Second, in terms of incentivizing effort, teachers working under P4P contracts elicited better performance from their students than did teachers working under FW contracts (holding advertised contracts and hence constant). The improvement in student achievement was 0.09 standard deviations per year on average across the two years. When we compare effects separately by program year, the effect is largest (0.16 standard deviations) in the second year of the study. These effects are large in relation to the shallow learning gradients we observe in Rwandan primary schools.

In addition to teacher characteristics and student outcomes, we observe a range of teacher behaviors. These behaviors corroborate our first finding: P4P recruits performed no worse than the FW recruits in terms of their presence, preparation, and observed pedagogy. They also indicate that the learning gains brought about by those experiencing P4P contracts may have been driven, at least in part, by improved teacher presence and pedagogy. Teacher presence was 6 percentage points higher among recruits who experienced the P4P contract compared to recruits who experienced the fixed-wage contract. (A sizeable impact given that baseline teacher presence was close to 90 percent.) And teachers who experienced P4P were more effective in their classroom practices than teachers who experienced FW by 0.26 points, as measured on a 4-point scale.

Third, in terms of retention, teachers working under P4P contracts were no more likely to quit during the two years of the experiment than teachers working under FW contracts. There was also no evidence of differential selection-out on baseline teacher characteristics by experienced contract. Teachers with a low baseline skill score (measured using a grading task) were significantly more likely to quit than teachers with a high baseline skill score. However, this selection-out was not more pronounced among the group working under P4P than among those working under FW. Teachers with a high baseline intrinsic motivation score (measured using a framed Dictator Game) were significantly more likely to quit than teachers with a low baseline intrinsic motivation score but, again, this selection-out was not more pronounced among the group working under P4P than among those working under FW. On the retention margin, we therefore find little evidence to support claims made by either proponents or opponents of pay-for-performance.

To sum up, we find that the recruitment, effort, and retention effects of P4P combine to raise learning quality. In the second year of the study, we estimate the total effect of P4P to be 0.21 standard deviations of pupil learning. One quarter of this impact can be attributed to selection at the recruitment stage, with the remaining three-quarters arising from increased effort on-the-job, including along incentivized dimensions such as teacher presence and observed pedagogy.

This paper makes several contributions. First, our results add to a small empirical literature on the recruitment of civil servants in low and middle-income countries. Existing papers have examined the impact of advertising higher base salaries (Dal Bó et al. (2013) for civil servants in Mexico; Deserranno (2019) for village promoters in Uganda) and better career track opportunities (Ashraf et al. (2016) for community health workers in Zambia) and found mixed results. In Mexico and Zambia, the interventions attracted both skilled and motivated workers, while in Zambia the offer of higher pay made it harder to recruit the most socially motivated agents, although overall applications increased. We study a different, and much debated, compensation policy: pay-for-performance. We find no impact of P4P on applications or on the measured skill of recruits at baseline and a negative impact of P4P on the measured intrinsic motivation of recruits at baseline. However, when we use our two-tiered experiment to isolate any recruitment effect in subsequent performance on-the-job, we find that these more 'money-motivated' recruits did not perform any worse in the classroom, if anything the reverse.

There is a large literature studying the impact of compensation policies on the composition of the teaching workforce in high-income setting such as the US. Several papers have examined IMPACT, the teacher evaluation system introduced in District of Columbia public schools, whereby financial incentives are linked to multiple measures of teacher performance (including student test scores). Dee and Wyckoff (2015) use a regression discontinuity design to show that low-performing teachers were more likely to quit voluntarily, while Adnot et al. (2017) subsequently confirm that these 'quitters' were replaced by higher-performers. Biasi (2019) studies a different reform: Act 10, passed in Wisconsin in 2011, that prompted approximately half of the State's school districts to replace seniority-based pay schedules with flexible salary schemes that allow pay to vary with performance. Consistent with predictions from a simple Roy model, Biasi finds that high-quality teachers (measured using pre-Act 10 teacher value-added) were more likely to move to districts with flexible pay, and were less likely to quit, than low-quality teachers. To the best of our knowledge,

there has not yet been a prospective, experimental study of pay-for-performance on the composition of the teaching workforce in a high-income setting. While there are obviously important political and economic differences between such settings and our context in Rwanda, key features of the labor market for teachers are similar (c.f. the significant fraction of trained teachers who are employed in non-teaching jobs and who, on average, draw higher average salaries than their teaching peers).[1] As such, our methodology and results also contributes to this broader literature.

Second, on the effort margin, we add to growing evidence that performance contracts—and in particular, individual-level schemes based on a 'growth percentile' student learning metric of the form proposed by Barlevy and Neal (2012)—can incentivise teachers to deliver improved learning outcomes (Loyalka et al. (2019) in China; Gilligan et al. (2018) in Uganda; Mbiti et al. (2018) in Tanzania). We add to these papers by: (i) studying the impact of a *composite* performance metric that places equal weight on the Barlevy-Neal percentile rank and three measures of teachers' inputs into the classroom (presence, lesson planning, and observed pedagogy), and (ii) comparing the effectiveness of contracts, P4P versus FW, that are *budget neutral* in salary. Both innovations were politically salient to our local partner, the Rwanda Education Board.

Third, a methodological contribution of the paper—in addition to the experimental design—is the way in which we develop a pre-analysis plan. In our registered plan (AEARCTR-0002565), we pose three questions. What outcomes to study? What hypotheses to test for each outcome? And how to test each hypothesis? We answered the 'what' questions on the basis of theory (in the form of a simple model), political relevance and available data. With these questions settled, we then answered the 'how' question using blinded data. Specifically, we used a blinded dataset that allowed us to learn about a subset of the statistical properties of our data without deriving hypotheses from realized treatment responses, as advocated by, e.g., Olken (2015).[2] This approach achieves power gains by choosing from among specifications and test statistics on the basis of simulated power, while protecting against the risk of false positives that could arise if specifications were chosen specifically on the basis of their realized statistical significance. The spirit of this approach is similar to recent

[1] The 2017 Rwanda Labour Force Survey includes a small sample of recent teacher training college graduates (i.e. below age 30). Of these, 37 percent were in teaching jobs earning an average salary of 43,431 RWF but 15 percent were in non-teaching jobs earning a higher average salary of 56,347 RWF, a premium of close to 30 percent.

[2] Although we have not found prior examples of such blinding in economics, Humphreys et al. (2013) argue for and undertake a related approach with partial endline data in a political science application.

work by Anderson and Magruder (2017) and Fafchamps and Labonne (2017).[3] For an experimental study in which one important dimension of variation occurs at the labor-market level—and so is potentially limited in power—the gains from these specification choices are particularly important. The results reported in our pre-analysis plan demonstrate that, with specifications appropriately chosen, the study design is well powered, such that even null effects would be of both policy and academic interest.

The remainder of the paper is organized as follows. Section 2 sets out the study design in greater detail. Section 3 provides a detailed description of the data. Sections 4 and 5 report results, and Section 6 concludes.

2 Experimental design

The study took place during the actual recruitment of civil service teaching jobs in upper primary in six districts of Rwanda in 2016.[4] The design draws on the 'surprise' two-stage randomizations of Karlan and Zinman (2009), Ashraf et al. (2010), and Cohen and Dupas (2010) in credit-market and public-health contexts. Both tiers of this experiment are built around the comparison of two contracts regarding a bonus payment, on top of existing teacher salaries, and managed by Innovations for Poverty Action in coordination with the Rwanda Education Board (REB). The first of these contracts is a pay-for-performance contract, which pays RWF 100,000 (approximately 15 percent of annual salary) to the top 20 percent of upper-primary teachers within a district, as measured by a composite performance metric briefly described in the Introduction and detailed in the pre-analysis plan. The second is a fixed wage contract that provides RWF 20,000 to all upper-primary teachers. As we discuss in detail in Section 2.2 below, the second-stage (surprise) randomization implies that applicants may experience a different contract from the advertised one to which they applied.

This design gives rise to four distinct types of recruits placed in schools, as summarized in Figure 1. *Potential applicants*—not all of whom are observed—are assigned to either advertised FW or advertised P4P contracts, depending on the labor market in which they reside. Those

[3] In contrast to those two papers, we forsake the opportunity to undertake exploratory analysis because our primary hypotheses were determined *a priori* by theory and policy relevance. In return, we avoid having to discard part of our sample, with associated power loss.

[4] In Rwanda, upper primary refers to grades 4, 5, and 6; schools themselves typically include grades 1 through 6.

who actually apply, and are placed into schools, fall into one of four groups. For example, group 'a' in Figure 1 denotes teachers who applied to jobs advertised as FW, and who were placed in schools assigned to FW contracts, while group 'c' denotes teachers who applied to jobs advertised as FW and were then placed in schools re-randomized to P4P contracts. The key insight of such a surprise re-randomization is that comparisons between groups a and b, and between groups c and d, allow us to learn about a 'pure' compositional effect of pay-for-performance contracts on teacher performance in the school, whereas comparisons along the diagonal of a–d are informative about the total effect of such contracts, along both extensive and intensive margins.

Figure 1: Treatment groups among recruits placed in study schools

		Advertised	
		FW	P4P
Experienced	FW	a	b
	P4P	c	d

The academic year in Rwanda runs from February–November, with new hires typically recruited between November and January. The timeline for the study was therefore as follows. In November 2015, as soon as districts revealed the positions to be filled, we announced the advertised contract assignment. In addition to radio, poster, and flyer advertisements, and the presence of a person to explain the advertised contracts at District Education Offices, we also held three job fairs at central locations to promote the interventions. These job fairs were advertised through WhatsApp networks of Teacher Training College graduates. Applications were then submitted in December. In January 2016, all districts held screening examinations for potential candidates. Successful candidates were placed into schools during February–March. We enrolled schools into the study on a rolling basis as they received recruits and allocated them to teaching positions in upper-primary grades. Our baseline survey was conducted in March 2016. Schools were assigned to treatments immediately following the baseline survey. We then measured teacher inputs over the course of the 2016 and 2017 school years, and measured learning outcomes at the end of each of the two academic years.

2.1 First-tier randomization: Advertised contracts

Our aim in the first tier was to randomize distinct labour markets to contracts, since this would 'treat' all potential applicants in a given labour market with a particular contract enabling us to assess any selection response. Discussions with REB during 2015 indicated that (i) few individuals apply for teaching jobs in multiple districts, and (ii) individuals are eligible for jobs defined by their subject specialization (there are five subjects: math, science, English, Kinyarwanda, and social studies). Accordingly, in November 2015 we defined a labour market in terms of a district-by-subject pair and randomly assigned treatment across 30 pairs (6 districts x 5 subjects).[5] All new primary posts within a P4P district-by-subject pair were to be advertised with a P4P contract, and all new primary posts within a FW district-by-subject pair were to be advertised with a FW contract.

In January 2016, we discovered that districts actually solicited applications at the slightly coarser district-by-*subject-family* level, aggregating subjects into three subject families that correspond to the degree types issued by Teacher Training Colleges: math and science (TMS); modern languages (TML); and social studies (TSS). We have 18 such labor markets defined by the product of district and subject family. The result of the randomized assignement is that 7 labor markets can be thought of as being in a 'P4P only' advertised treatment (modern language teaching in Gatsibo and Kirehe, math and science teaching in Kayonza and Nygatore, and social studies teaching in Ngoma, Nygatore, and Rwamagama); 7 in a 'FW only' advertised treatment (modern language teaching in Kayonza and Rwamagama, math and science teaching in Kirehe and Ngoma, and social studies teaching in Gatsibo, Kayonza, and Kirehe); and 4 in a 'Mixed' advertised treatment (modern language teaching in Ngoma and Nygatore, and math and science teaching in Gatsibo and Rwamagama). To illustrate this Mixed treatment, an individual living in Ngoma with a qualification to teach modern languages could have applied to the modern languages pool, in which case they would have been eligible for either an advertised post in English on a FW contract, or an advertised post in Kinyarwanda on a P4P contract. In contrast, someone living in Gatsibo with a qualification to teach modern languages would have been subject to the 'P4P only' treatment; he/she could have applied for either an English or Kinyarwanda post, but both would have been

[5]This randomization was performed in MATLAB by the PIs.

on a P4P contract. Empirically, we will consider the Mixed treatment as a separate arm. We will estimate a corresponding advertisement effect but interpret this only as an incidental parameter.

This first-tier randomization was accompanied by an advertising campaign to increase awareness of the new posts and their associated contracts, including organization of job fairs at Teacher Training Colleges. As we discuss in Section 3 below, extensive data on potential applicants were collected at these job fairs. Advertisements also took place over the radio, in person at District Education Offices, and through dissemination of printed materials in capitals of the study districts. These advertisements emphasized that the contracts were available for recruits placed in the 2016 school year and that the payments would continue into the 2017 school year.

2.2 Second-tier randomization: Experienced contracts

Our aim in the second tier was to randomize the schools to which REB had allocated the new posts to contracts. A school was included in the sample if it had at least one new post that was filled *and* assigned to an upper-primary grade (grades 4, 5 and 6, hereafter P4, P5, and P6). Following a full baseline survey in February 2016, sample schools were randomly assigned to either P4P or Fixed Wage. Of the 164 schools in the second tier of the experiment, 85 were assigned to P4P and 79 were assigned to Fixed Wage contracts.

All upper-primary teachers within each school received the new contract. At individual applicant level, this amounted to re-randomization and hence a change to the initial assignment for some new recruits. A natural concern is that individuals who applied under one contract, but who were eventually offered another contract, might have experienced disappointment (or other negative feelings) which then had a causal impact on their behaviour. To mitigate this concern, all new recruits were offered an *end-of-year retention bonus* of RWF 80,000 on top of their school-randomized P4P or FW contract. An individual who applied under advertised P4P in the hope of receiving RWF 100,000 from the scheme, but who was subsequently re-randomized to experienced FW, was therefore still eligible to receive RWF 100,000 (RWF 20,000 from the FW contract plus RWF 80,000 as a retention bonus). Conversely, an individual who applied under advertised FW safe in the knowledge of receiving RWF 20,000 from the scheme, but who was subsequently re-randomized to experienced P4P, was still eligible for at least RWF 80,000. None of the recruits objected to the (re)randomization or turned down their re-randomized contract.

Of course, disappointment effects may still be present in on-the-job performance. When testing hypotheses relating to teacher value-added, we include a secondary specification with an interaction term to allow the estimated impact of experienced P4P to differ by advertised treatment. Intuitively, in terms of Figure 1, we compare groups a and c to obtain the impact for the advertised FW subgroup, and then groups b and d to obtain the impact for the advertised P4P subgroup. If disappointment effects are an issue, then since it is groups b and c who receive the 'surprise', one would expect the impact for the advertised P4P subgroup to be larger than the impact for the advertised FW subgroup. In fact, we find the reverse, although the interaction term is not significant at conventional levels.

In addition to information about the end-of-year retention bonus, teachers in P4P schools were also told that the 2016 performance award—determined by multiple teacher-input observations as well as beginning- and end-of-year student assessments—was conditional on remaining in post during the 2016 school year, and would be paid early in 2017. The experiment continued in the same 164 schools for the 2017 school year. Schools were contacted by telephone in February 2017 to remind them of the continuation of the scheme. Teachers in FW schools were told that they would receive the RWF 20,000 award, and teachers in P4P schools were told that the 2017 performance award—calculated in a similar fashion to the 2016 award—would be paid early in 2018. Our enumerators stressed that both payments were conditional on remaining in post during the 2017 school year.

3 Data

The primary analyses make use of several distinct types of data. Conceptually, these trace out the causal chain from the advertisement intervention to a sequence of outcomes: that is, from the candidate's application decision, to the set (and attributes) of candidates hired into schools, to the learning outcomes that they deliver, and, finally, to the teacher's decisions to remain in the schools. In this section, we describe the administrative, survey, and assessment data available for each of these steps in the causal chain. A schematic of these data sources, and their timing in relation to intervention, is laid out in Appendix Figure A.1. Our understanding of these data informs our choices of specification for analysis, as discussed in detail in the pre-analysis plan.

3.1 Applications

Table 1 summarizes the applications for the newly advertised jobs, submitted in January 2016, across the six districts.[6] Of the 2,185 applications in total, 1,963 come from candidates with a Teacher Training College (TTC) degree—we term these *qualified* since, at least in principle, a TTC degree is required for the placements at stake. In addition to submitting their TTC qualifications, applicants were required to undertake a district-level exam in order to be considered for a post. This final step was added only after applications were submitted, in a change of regulation from the Rwanda Education Board. Each district constructed its own assessment for this purpose, and districts used the same assessment tool across all subjects of application. We refer to applications from individuals who hold both a TTC degree and who sat a district-specific exam as *complete*.[7] In the table, we present TTC scores, genders, and ages—the other observed CV characteristics—for all qualified applicants, regardless of whether their application includes a district exam score.

Table 1: Application characteristics, by district

	Gatsibo	Kayonza	Kirehe	Ngoma	Nyagatare	Rwamagana	All
Applicants	390	310	462	381	327	315	2,185
Qualified	333	258	458	365	272	277	1,963
Has TTC score	317	233	405	338	260	163	1,716
Mean TTC score	0.53	0.54	0.50	0.53	0.54	0.55	0.53
SD TTC score	0.14	0.15	0.19	0.15	0.14	0.12	0.15
District score	273.00	198.00	312.00	173.00	177.00	78.00	1,211.00
Qualified female	0.53	0.47	0.45	0.50	0.44	0.45	0.48
Qualified age	27.32	27.78	27.23	27.23	26.98	27.50	27.33

The 2,185 applications come from 1,424 unique individuals, of whom 1,194 have a TTC qualification. Qualified applicants complete an average of 1.61 applications in study districts, with 62 percent of qualified applicants completing only one application.

3.2 Teacher attributes

Following the application stage, successful applicants to posts in study districts were placed into schools by District Education Officers, and were assigned to particular grades, subjects, and streams by their head teachers. A primary school in a study district was enrolled in the study if two

[6] These data were obtained from the six district offices and represent a census of applications for the new posts.
[7] A small number of candidates sat for district exams despite not having a TTC degree.

conditions were met: the District Education Officer placed at least one new recruit in this school *and* the head teacher assigned at least one of these new recruits to an upper-primary teaching role. Note that once a school had been enrolled, any teacher—placed recruit or incumbent—who was assigned to teach one of the five 'core curricular' subjects in upper-primary grades 4, 5 or 6 was eligible for the intervention.

We visited the enrolled schools at baseline in February 2016, and collected data using three broad types of instruments: school surveys, teacher surveys, and teacher 'lab' measures. We describe these measures below and in Table 2. In doing so, we summarize the attributes of three mutually exclusive types of teachers: recruits, who were hired by the district, incumbents, who were teaching in the school the previous year, and 'other' teachers, who include new informal and community hires not affected by the first-tier randomization.

School surveys. These were administered to head teachers, or their deputies, at baseline and included a variety of data on management practices—not documented here—as well as administrative records of teacher attributes, including age, gender, and qualifications. The data cover all teachers in the school, regardless of whether they were eligible for the intervention.

Teacher surveys. These were administered to all teachers responsible for at least one upper-primary, core-curricular subject and included questions about demographics, household background, training, qualifications and experience, and earnings. Of particular note are attributes that might be associated with both the likelihood of selecting into P4P contracts relative to FW contracts and with subsequent performance: the 'Big-5' personality traits, self esteem, and the locus of control (Almlund et al., 2011; Callen et al., 2018; Dal Bó et al., 2013; Donato et al., 2017; Gensowski, 2017; John, 1990).

'Lab-in-the-field' instruments. We used a series of incentivized 'lab-in-the-field' tasks to provide additional measures of teacher attributes.

In a framed version of the *Dictator Game*, teachers were given 2,000 Rwandan Francs (RWF) and asked how much of this money they wished to allocate towards providing school supply packets to students in their schools, and how much they wished to keep for themselves. Each packet contained one notebook and pen and was worth 200 RWF. They could decide to allocate any amount, from

Table 2: Baseline teacher characteristics

	Recruit	Incumbent	Other
Characteristics from school survey (all teachers)			
Female	0.40	0.48	0.46
	(0.49)	(0.50)	(0.50)
Age	26.34	35.40	35.17
	(4.41)	(8.98)	(8.65)
Observations	329	2,854	221
Characteristics from teacher survey (upper primary teachers only)			
Big 5 personality traits			
Conscientiousness	6.07	6.01	6.03
	(0.42)	(0.55)	(0.57)
Extraversion	4.83	4.73	4.30
	(1.02)	(10.31)	(0.97)
Agreeableness	5.69	5.87	5.85
	(0.76)	(0.70)	(0.67)
Openness to experience	5.31	5.07	5.36
	(0.82)	(1.03)	(0.78)
Neuroticism	1.92	1.60	1.35
	(1.22)	(1.08)	(1.00)
Big Five index	-0.03	0.00	0.10
	(0.46)	(0.58)	(0.41)
Locus of control (Rotter)	3.06	3.00	2.63
	(0.57)	(0.71)	(0.72)
Self esteem (Rosenberg)	29.19	30.50	29.67
	(3.23)	(2.12)	(3.83)
Observations	251	1,067	78
Lab measures (upper primary teachers only)			
Dictator game: share sent	0.27	0.43	0.46
	(0.32)	(0.35)	(0.34)
Share choosing lottery...			
A	0.36	0.29	0.31
B	0.18	0.18	0.12
C	0.16	0.16	0.14
D	0.12	0.15	0.10
E	0.19	0.22	0.33
Grading task score (IRT)	-0.16	0.04	-0.01
	(0.89)	(0.90)	(0.95)
Competition game: share choosing to compete	0.65	0.66	0.73
Observations	250	1,066	78

Note: Means with standard deviations in parentheses. Total observations reported separately for each data source.

zero to all 2,000 RWF, which would supply ten randomly chosen students with a packet. The purpose of this game was to measure teachers' other regarding preferences towards students and is a refinement of approaches used elsewhere.[8] Table 2 shows that recruits gave on average 27 percent of the stake to the schools' students—substantially less than the average donated share of 43 percent by incumbent teachers.

Next, teachers participated in a *Lottery Choice* task designed to measure their degree of risk aversion. Based on existing instruments (Binswanger, 1980; Chetan et al., 2010; Eckel and Grossman, 2008), this task asks participants to choose between five lotteries, which include one certain outcome (here, labelled as Option A) and a series of alternatives increasing in their returns, but also their riskiness. Table 2 shows that more than a third of recruits chose the certain outcome, but beyond this, choices are fairly evenly spread across the remaining alternatives.

Teachers also undertook a *Grading Task* which measured their mastery of the curriculum in the primary subject that they teach (c.f. similar tasks used to by Bold et al. (2017) to construct the World Bank's Service Delivery Indicators). Teachers were asked to grade a student examination script, and had 5 minutes to determine if a series of student answers were correct or incorrect. They received a fixed payment for participation that did not depend on performance.

This grading task also served as the first stage in a *Competition Game*, based on Niederle and Vesterlund (2007), which has been shown elsewhere to be associated with gender differences in the taste for competitive careers (Buser et al., 2014). Following the fixed-pay grading task that provides our measure of teacher skill, teachers were asked to undertake a second grading exercise. This second grading task took the form of a tournament: only the top 20 percent of teachers within a given subject and district would receive a payout, and the payout would be 5 times the payout for the fixed-pay grading task.[9] Finally, in a third round teachers were allowed to choose between the fixed-pay and tournament payment schemes. Table 2 highlights that nearly two thirds

[8]Lagarde and Blaauw (2014) use a Dictator Game with patients as the second player to measure nurses' other regarding preferences in South Africa. Brock et al. (2016) measure health workers' pro-social behaviour employing a Dictator Game where the recipient is an (anonymous) non-clinician versus clinician to compare motivations to patients versus peers in Tanzania. Other studies have used a named organization as the recipient, following early lab experimental work (Eckel and Grossman, 1996). See, e.g., the Indonesian Red Cross in Sheheryar and Keefer (2016), a locally well-known NGO that provides care to HIV/AIDS patients in Ashraf et al. (2014), and an existing local public health NGO Deserranno (2019).

[9]This payout structure is modified from the original Competition Game of Niederle and Vesterlund (2007), who compare a piece rate with a tournament in which the winner receives a multiple of that same piece rate. We made this change to mirror the contractual choice facing applicants in our study.

of both recruits and incumbents chose the tournament scheme; this decision (residualized to account for differences in actual ability) will provide a measure of the 'taste for competition' used in the secondary analysis described in the pre-analysis plan.

3.3 Student learning

Student learning was measured via assessments taken at the start and end of the 2016 school year, and the end of the 2017 school year (indexed by {0,1,2}, respectively, in subsequent notation). These student assessments play a dual role in our study: they provide the primary measure of learning for analysis of program impacts, and they were used in the evaluation of teachers in the experienced P4P arm for purposes of performance awards, as discussed in the pre-analysis plan.[10]

We developed comprehensive subject- and grade-specific, competency-based assessments for grades 4, 5 and 6.[11] These assessments were based on the new Rwanda national curriculum and covered the five core subjects: Kinyarwanda, English, Mathematics, Sciences, and Social Studies. We developed one assessment per grade-subject, with students at the beginning of the year being assessed on the prior year's material (and a special grade 3 assessment developed for the purpose of assessing grade 4 students at the beginning of the year). Each test aimed to cover the entire curriculum for the corresponding subject and year, with questions becoming progressively more difficult as a student advanced in the test. The questions were a combination of multiple choice and fill-in diagrams.[12]

In each round, we randomly sampled a subset of students from each grade to take the test. In Year 1 of the study, both baseline and endline student samples were drawn from the official school register of enrolled students (complied by the school at the beginning of the year). This was done to ensure that the sampling protocol did not create incentives for strategic exclusion of students. In Year 2, students were assessed at the end of the year only, and were sampled from a listing that we collected in the second trimester.

[10] As a robustness check, we will also test for impacts of experienced P4P on scores from national exams taken in grade 6. These scores were not included in the incentive metric and will therefore enable us to check for the possibility of 'teaching to the test'.

[11] The tests were developed in cooperation with local and international experts, and in consultation with the Ministry of Education. They were extensively piloted and revised during and after piloting.

[12] In piloting, all student tests were administered in English but we found that grade 4 students had not yet received the level of English instruction necessary to be adequately measured using an English-based exam. Grade 4 tests were therefore translated and administered in Kinyarwanda throughout the study.

Student samples were stratified by teaching *streams* (subgroups of students taught together for all subjects). In Round 0, we sampled a minimum of 5 pupils per stream, and oversampled streams taught in at least one subject by a new recruit to fill available spaces, up to a maximum of 20 pupils per stream and 40 per grade. In rare cases of grades with more than 8 streams, we sampled 5 pupils from all streams. In Round 1, we sampled 10 pupils from each stream: 5 pupils retained from the baseline (if the stream was sampled at baseline) and 5 randomly sampled new pupils. We included the new students to alleviate concerns that teachers in P4P schools might teach (only) to previously sampled students. In Round 2, we randomly sampled 10 pupils from each stream using the listing for that year.[13] Resulting sample sizes are presented in Table 3.

Table 3: Pupil and assessment descriptive statistics

	Round		
	Baseline	Round 1	Round 2
Schools	.	.	.
Streams	1,629	.	1,772
Total upper-primary pupils	67,371	.	72,412
Pupils sampled for test	14,672	16,067	17,722
Pupils taking exam	13,831	14,310	16,874
Student-subjects assessed	69,141	71,550	84,370
E[z]	0.00	0.00	0.00
Var[z]	0.83	0.78	0.78

Note: Enrollment figures taken from official pupil registration data, updated annually, and hence not collected in Round 1 at the end of Year 1.

The tests were orally administered by trained enumerators. Students listened to an enumerator as he/she read through the instructions and test questions, prompting students to answer. The exam was timed for 50 minutes, allowing for 10 minutes per section. Enumerators administered the exam using a timed proctoring video on electronic tablets.[14] Individual student test results were kept confidential from teachers, parents, head teachers, and Ministry of Education officials, and have only been used for performance award and evaluation purposes in this study.

Responses were used to estimate a measure of student learning (for a given student in a given round and given subject in a given grade) based on a *two-parameter Item Response Theory (IRT)*

[13]Consequently, the number of pupils assessed in Year 2 who have also been assessed in Year 1 (either at baseline or endline) is limited. Because streams are reshuffled across years and because we were not able to match Year 2 pupil registers to Year 1 registers in advance of the assessment, it was not possible to sample pupils to maintain a panel across years while continuing to stratify by stream.

[14]The proctoring videos were an additional safeguard to ensure consistency in test administration and timing.

model, which was estimated using Stata's `irt 2pl` command. We use empirical Bayes estimates of student ability from this model as our measure of a student's learning level in a particular grade.

3.4 Teacher inputs

We collected data on teachers' inputs into the classroom. This was undertaken in P4P schools only during Year 1, and in both P4P and FW schools in Year 2. These measures contribute to the incentivized teacher performance metric in P4P schools, as described in the pre-analysis plan. This composite metric is based on three input measures (teacher presence, lesson preparation and pedagogical practice), and one output measure (student performance)—the '4Ps'. Here we describe the input components measured.

To assess the three inputs, P4P schools received three unannounced surprise visits: two spot checks during Summer 2016, and one spot check in Summer 2017. During these visits, Sector Education Officers (SEOs) from the District Education Offices (in Year 1) or IPA staff (for logistical reasons, in Year 2) observed teachers and monitored their presence, preparation and pedagogy with the aid of specially designed tools.[15] FW schools also received an unannounced visit in Year 2, at the same time as the P4P schools. Table 4 shows summary statistics for each of these three input measures over the three rounds of the study. are presented in Table 5.

Presence is defined as the fraction of spot-check days that the teacher is present at the start of the school day. The SEOs recorded teacher presence after speaking with the head teacher at the start of the school day during each unannounced visit. In order for the SEO to record a teacher present, the head teacher had to physically show the SEO that the teacher was in school rather than relying on an attendance roster.

Lesson *preparation* is defined as the planning involved with daily lessons, and is measured through a review of teacher written weekly lesson plans. Prior to any spot checks, teachers in grades 4, 5, and 6 in P4P schools were shown how to fill out a lesson plan in accordance with REB

[15]Training of SEOs took place over two days. Day 1 consisted of an overview of the study and its objectives and focused on how to explain the intervention (in particular the 4Ps) to teachers in P4P schools. During Day 2, SEOs learned how to use the teacher monitoring tools and how to conduct unannounced school visits. SEOs practiced using these monitoring tools by viewing videos recorded during pilot visits. Training sessions were led by staff experienced in teacher evaluation to ensure that SEOs applied the rubrics consistently. SEOs were briefed on the importance of not informing teachers or head teachers ahead of the visits. Field staff monitored the SEOs adherence to protocol, including through random phone calls to head teachers.

Table 4: Measures of teacher inputs in P4P schools

	Mean	St Dev	Obs
Year 1, Round 1			
Teacher present	0.97	(0.18)	661
Has lesson plan	0.54	(0.50)	598
Classroom observation: Overall score	2.01	(0.40)	645
Lesson objective	2.00	(0.70)	645
Teaching activities	1.94	(0.47)	645
Use of assessment	1.98	(0.50)	643
Student engagement	2.12	(0.56)	645
Year 1, Round 2			
Teacher present	0.96	(0.21)	648
Has lesson plan	0.54	(0.50)	598
Classroom observation: Overall score	2.27	(0.41)	639
Lesson objective	2.21	(0.77)	638
Teaching activities	2.17	(0.46)	638
Use of assessment	2.23	(0.48)	638
Student engagement	2.46	(0.49)	639
Year 2, Round 1			
Teacher present	0.90	(0.31)	739
Has lesson plan	0.79	(0.41)	610
Classroom observation: Overall score	2.36	(0.35)	636
Lesson objective	2.47	(0.66)	636
Teaching activities	2.26	(0.44)	634
Use of assessment	2.25	(0.47)	635
Student engagement	2.48	(0.46)	636

Notes: Descriptive statistics presented are for upper-primary teachers only. Overall score for classroom observation is average of four components: Lesson objective, Teaching activities, Use of assessment, and Student engagement, with each component scored on a scale from zero to three.

guidance.[16] Specifically, SEOs visited schools and provided teachers with a template to help prepare three key components of a lesson—write out the lesson objective, list the instructional activities, and list the types of assessment that will be carried out. A 'hands-on' session then enabled teachers to practice writing lesson plans using this template before incorporating it in their daily teaching practice. During the SEO's unannounced visit, he/she collected the daily lesson plans (if any had been prepared) from each teacher. Field staff subsequently used a lesson planning scoring rubric to provide a subjective measure of quality. Because a substantial share of upper-primary teachers do not have a lesson plan on a randomly chosen audit day, we use the presence of such a lesson plan as a summary measure in both the incentivized contracts and as an outcome for anlaysis.

Pedagogy is defined as the practices and methods that teachers use in order to impact student learning. We collaborated with both the Ministry of Education and REB in May and June 2015 to develop a monitoring instrument to measure teacher pedagogy through classroom observation. Our classroom observation instrument measured objective teacher actions and skills as an input into scoring teachers' pedagogical performance, using a rubric adapted from the Danielson Framework for Teaching, which is widely used in the U.S. (Danielson, 2007). The observer evaluated the teachers' effective use of 21 different activities over the course of a full 45-minute lesson.[17] Based on these observations and a detailed rubric, the observer provided a subjective score, on a scale representing mastery from zero to three, of four components of the lesson: communication of lesson objectives, delivery of material, use of assessment, and student engagement.[18] The teacher's incentivized score, as well as our measure of pedagogy, is defined as the average of these ratings across the four domains.

3.5 Job fairs

Although not part of our core analysis, we will also report results using data collected at our TTC job fairs. At each of the three events that were held in December 2015—during the application period affected by the intervention—we invited attendees to participate in the same survey and

[16] To isolate the effects of performance pay, aspects of training were kept to a minimum and focused on how teachers could meet the targeted metrics.

[17] Though not structured as a strict time-on-task measure, this aspect is similar to the Stallings Observation System (Stallings et al., 2014).

[18] Similar rubric-based scoring has been used in other field experiments, including Glewwe et al. (2010) who measure teacher effort with a similar intensity scale in a teacher incentive study in Kenya.

'lab-in-the-field' tasks that were (subsequently) administered to teachers at baseline. We were then able to link responses to application and placement decisions, and (for the subset of attendees who became placed recruits) to the full set of study outcomes. These job fair data are useful because they provide an insight into the pool of *potential* applicants to FW and P4P positions. But, of course, this insight is only partial since participants 'selected in' to attend these informational events and are not necessarily representative of the wider pool.

Table 5: Job-fair participants

	Mean	St. Dev.	Observations
Survey characteristics			
Female	0.46	(0.50)	203
Age	23.59	(2.26)	202
Big 5 personality traits			
Conscientiousness	6.06	(0.63)	202
Extraversion	3.95	(0.65)	202
Agreeableness	5.97	(0.67)	202
Openness to experience	5.50	(0.83)	202
Neuroticism	5.20	(33.40)	202
Lab measures			
Dictator game: share sent	0.33	(0.32)	203
Share choosing lottery...			
A	0.34	(0.48)	203
B	0.15	(0.36)	203
C	0.14	(0.35)	203
D	0.12	(0.32)	203
E	0.25	(0.43)	203
Grading task, pct correct	0.32	(0.13)	156
Competition game: share choosing to compete	0.64	(0.48)	194

4 Results

We set out to address six questions, which each correspond to a primary hypothesis that we test, and a small number of associated secondary hypotheses that represent alternative measures or mechanisms. These hypotheses were specified in the pre-analysis plan, using the theoretical framework set out in Appendix A; they are the following:

I. Advertised P4P induces differential application qualities;

II. Advertised P4P affects the observable skills of recruits placed in schools;

III. Advertised P4P induces differentially 'intrinsically' motivated recruits to be placed in schools;

IV. Advertised P4P induces the *selection* of higher- (or lower-)performing teachers, as measured by the learning outcomes of their students;

V. Experienced P4P creates *incentives* which contribute to higher (or lower) teacher performance, as measured by the learning outcomes of their students;

VI. Selection and incentive effects are apparent in the composite 4P performance metric.

For each of these hypotheses, four questions determine how they are tested: (a) What outcome measure will be used? (b) On what sample will this be estimated? (c) What test statistic will be used? And (d) How will inference be undertaken on this test statistic? We summarize these design decisions in Table 6, copied from the pre-analysis plan, and provide details of each below.

Table 6: Summary of hypotheses, outcomes, samples, and specifications

Outcome	Sample	Test statistic	Randomization inference
Hypothesis I: Advertised P4P induces differential application qualities			
*TTC exam scores	Universe of applications	KS test of eq. (1)	$\mathcal{T}^A$
District exam scores	Universe of applications	KS test of eq. (1)	$\mathcal{T}^A$
TTC exam scores	Universe of applications	t_A in eq. (2)	$\mathcal{T}^A$
TTC exam scores	Applicants in the top $\hat{H}$ number of applicants, where $\hat{H}$ is the predicted number of hires based on subject and district, estimated off of FW applicant pools	t_A in eq. (2)	$\mathcal{T}^A$
TTC exam scores	Universe of application, weighted by probability of placement	t_A in eq. (2)	$\mathcal{T}^A$
Number of applicants	Universe of applications	t_A in eq. (3)	$\mathcal{T}^A$
Hypothesis II: Advertised P4P affects the observable skills of placed recruits in schools			
*Teacher skills assessment IRT model EB score	Placed recruits	t_A in eq. (11)	$\mathcal{T}^A$
Hypothesis III: Advertised P4P induces differentially 'intrinsically' motivated recruits to be placed in schools			
*Dictator-game donations	Placed recruits	t_A in eq. (5)	$\mathcal{T}^A$
Perry PSM instrument	Placed recruits retained through Year 2	t_A in eq. (5)	$\mathcal{T}^A$
Hypothesis IV: Advertised P4P induces the selection of higher-(or lower-) value-added teachers			
*Student assessments (IRT EB predictions)	Pooled Year 1 & Year 2 students	t_A in eq. (6)	$\mathcal{T}^A$
Student assessments	Pooled Year 1 & Year 2 students	t_A and t_{A+AE};	$\mathcal{T}^A$
		t_{AE} in eq. (7)	$\mathcal{T}^A \times \mathcal{T}^E$
Student assessments	Year 1 students	t_A in eq. (6)	$\mathcal{T}^A$
Student assessments	Year 2 students	t_A in eq. (6)	$\mathcal{T}^A$
Hypothesis V: Experienced P4P creates incentives which contribute to higher (or lower) teacher value-added			
*Student assessments (IRT EB predictions)	Pooled Year 1 & Year 2 students	t_E in eq. (6)	T^E
Student assessments	Pooled Year 1 & Year 2 students	t_E and t_{E+AE};	$\mathcal{T}^E$
		t_{AE} in eq. (7)	$\mathcal{T}^A \times \mathcal{T}^E$
Student assessments	Year 1 students	t_E in eq. (6)	$\mathcal{T}^E$
Student assessments	Year 2 students	t_E in eq. (6)	$\mathcal{T}^E$

Continues...

Table 6, continued

Outcome	Sample	Test statistic	Randomization inference
Hypothesis VI: Selection and incentive effects are apparent in the 4P performance metric			
*Composite 4P metric	Teachers, pooled Year 1 (experienced P4P only) & Year 2	t_A in eq. (8)	$\mathcal{T}^A$
Composite 4P metric	Teachers, pooled Year 1 (experienced P4P only) & Year 2	t_A and t_{A+AE};	$\mathcal{T}^A$
		t_E and t_{E+AE};	$\mathcal{T}^E$
		t_{AE} in interacted eq.	$\mathcal{T}^A \times \mathcal{T}^E$
Barlevy-Neal rank	As above		
Teacher attendance	As above		
Classroom observation	As above		
Lesson plan (indicator)	As above		

Primary tests of each family of hypotheses appear first, preceded by a superscript *; those that appear subsequently under each family without the superscript * are secondary hypotheses. Under inference, $\mathcal{T}^A$ refers to randomization inference involving the permutation of the *advertised* contractual status of the recruit *only*; $\mathcal{T}^E$ refers to randomization inference that includes the permutation of the *experienced* contractual status of the school; $\mathcal{T}^A \times \mathcal{T}^E$ indicates that randomization inference will permute both treatment vectors to determine a distribution for the relevant test statistic. Test statistic is a studentized coefficient or studentized sum of coefficients (a t statistic), except where otherwise noted (as in Hypothesis I); in linear mixed effects estimates of equation (6) and (7), which are estimated by maximum likelihood, this is a z rather than t statistic, but we maintain notation to avoid confusion with the test score outcome, z_{jbksr}.

4.1 Hypothesis I. Advertised P4P induces differential application qualities.

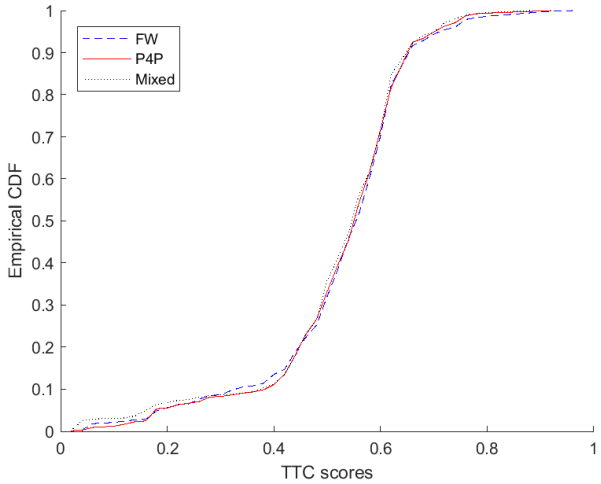
Primary test The primary test of this hypothesis is a non-parametric Kolmogorov-Smirnov (KS) test for a difference in distributions of applicant teacher training college final exam score across the advertised P4P and advertised FW labor markets. We can write the KS test-statistic as

$$T^{KS} = \sup_{y} \left| \hat{F}_{P4P}(y) - \hat{F}_{FW}(y) \right| = \max_{i=1,\ldots,N} \left| \hat{F}_{P4P}(y_i) - \hat{F}_{FW}(y_i) \right|. \tag{1}$$

Here, $\hat{F}_{P4P}(y)$ denotes the empirical cumulative distribution function of TTC score among applicants who applied under advertised P4P, evaluated at some specific TTC score y. Likewise, $\hat{F}_{FW}(y)$ denotes the empirical cumulative distribution function of TTC score among applicants who applied under advertised FW, evaluated at the same TTC score y.

We test the statistical significance of this difference in distributions by randomization inference. To do so, we repeatedly sample from the set of potential (advertised) treatment assignments $\mathcal{T}^A$ and, for each such permutation, calculate the KS test-statistic. The relevant p-value is then given by the share of such test statistics larger in absolute value than the test statistic estimated from the actual assignment.

Figure 2: Distribution of applicant TTC score, by advertised treatment arm



Consistent with the visual evidence in Figure 2, distributions of applicant TTC scores are statistically indistinguishable between the P4P and FW advertisement arms. The KS test-statistic has a value of 0.0264, with a p-value of 0.96. Randomization inference is well-powered, meaning

that we can rule out even small effects on the TTC score distribution: a 95 percent confidence interval based on inversion of the RI test rules out additive treatment effects outside of the range $[-0.02, 0.02]$. We therefore conclude that there was no meaningful impact of advertised P4P on the TTC final exam score among applicants.

Secondary tests For secondary tests of this hypothesis, we estimate a series of *weighted* regressions of the form

$$y_{iqd} = \tau_A T^A_{qd} + \gamma_q + \delta_d + e_{iqd}, \quad \text{with weights } w_{iqd} \tag{2}$$

where y_{iqd} denotes the TTC exam score of applicant teacher i with qualification q in district d. Treatment T^A_{qd} denotes the contractual condition under which a candidate applied.[19]

We focus on the impacts of advertised P4P under three specific selection rules:

- Impacts on the average quality, as measured by TTC score, of all applicants. This corresponds to $w_{iqd} = 1$ for all teachers.
- Impacts on the average ability of the top $\hat{H}$ applicants, where $\hat{H}$ is the predicted number hired in that district and subject based on outcomes in advertised FW district-subjects. This corresponds to weights $w_{iqd} = 1$ for the top $\hat{H}$ teachers in their application pool, and zero otherwise. The fraction hired is predicted from a regression of the number of actual hires on district and subject indicators, using FW applicant pools only.
- Impacts on applicants, weighted by their probability of hiring, using the FW district hiring probability. This corresponds to weights $w_{iqd} = \hat{p}_{iqd}$, where $\hat{p}_{iqd}$ is the estimated probability of being hired as a function of district and subject indicators, as well as a fifth-order polynomial in TTC exam scores, estimated using FW applicant pools only.

The first of these can be thought of as representing the consequences of advertised P4P for placed teacher quality under a random hiring rule; the second represents the outcome of advertised P4P under meritocratic selection on the basis of TTC exam scores alone; and the third represents

[19] Here and throughout the empirical specifications, we will define T^A_{qd} as a *vector* that includes indicators for both the P4P and mixed-treatment advertisement condition. However, for hypothesis testing, we are interested only in the coefficient on the pure P4P treatment. Defining treatment in this way ensures that only candidates who applied (and in subsequent sections, were placed) under the pure FW treatment are considered as the omitted category here, to which P4P recruits will be compared.

the consequences of advertised P4P under the status quo mapping from TTC scores to hiring probabilities.

The weighted regression parameter τ_A estimates the difference in (weighted) mean applicant skill induced by advertised P4P. To undertake inference about this difference in means, we use randomization inference, sampling repeatedly from the set of potential (advertised) treatment assignments $\mathcal{T}^{\mathcal{A}}$. Following Chung and Romano (2013), we studentize this parameter by dividing it by its (cluster-robust, clustered at the district-subject level) standard error to control the asymptotic rejection probability against the null hypothesis of equality of means. These are two-sided tests.[20] The absolute value of the resulting test statistic, $|t_A|$, is compared to its randomization distribution in order to provide a test of the hypothesis that $\tau_A = 0$.

In the simplest case, where all observations are weighted equally, our estimate of τ_A is -0.00. The studentized coefficient has a standard deviation of 0.011 under the sharp null. The randomization inference p-value is 0.95, indicating that we cannot reject the sharp null of no impact of advertised P4P.

We complete our secondary analysis of Hypothesis I by testing for differences in the number of applicants by treatment status, conditional on district and subject-family fixed indicators. We do so with a specification of the form

$$N_{qd} = \tau_A T^A_{qd} + \gamma_q + \delta_d + e_{qd}, \tag{3}$$

where q indexes subject families and d indexes districts; N_{qd} measures the number of qualified applicants in each district.[21] As above, we obtain the studentized test statistic t_A by dividing the estimated coefficient τ_A by the analytical estimate of its cluster-robust standard error, and use this t-statistic in our randomization inference.

In this application volume regression, our estimate of τ_A is -1.14. This has a RI p-value of 0.95, and a confidence interval of [-36.5, 44.1]. Thus, we fail to reject the null of no impact of advertised P4P on application volumes, though this is not as precisely estimated as the primary outcome.

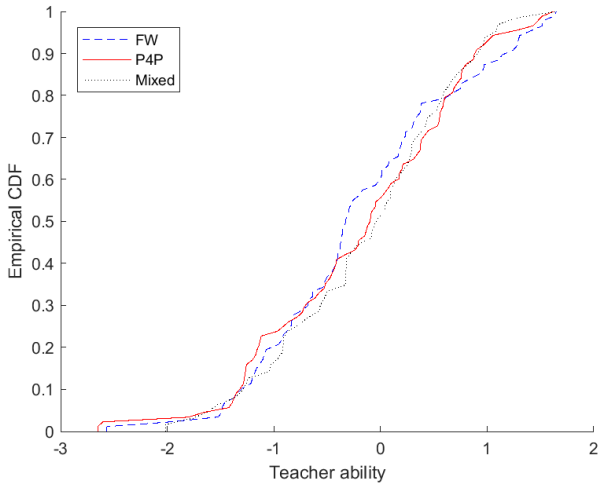
[20] We calculated p-values for two-sided tests as provided in Rosenbaum (2010) and in the 'Standard Operating Procedures' of Donald Green's Lab at Columbia (Lin et al., 2016).

[21] 'Qualified' here means that the applicant has a TTC degree. In addition to being a useful filter for policy-relevant applications, since only qualified applicants can be hired, in some districts' administrative data this is also necessary in order to determine the subject-family under which an individual has applied.

4.2 Hypothesis II. Advertised P4P affects the observable skills of placed recruits in schools.

Our primary (and only) test of this hypothesis uses our baseline estimate of skill among placed recruits. Specifically, we use empirical Bayes predictions from an IRT model of teacher skill in each subject,[22] which we denote by z_{iqd} for teacher i with qualification q in district d. Figure 3 plots the distribution of this measure by advertised treatment arm.

Figure 3: Distribution of placed teacher ability, by advertised treatment arm



To test the sharp null of no effects we estimate a regression of the form

$$z_{iqd} = \tau_A T^A_{qd} + \gamma_q + \delta_d + e_{iqd}. \tag{4}$$

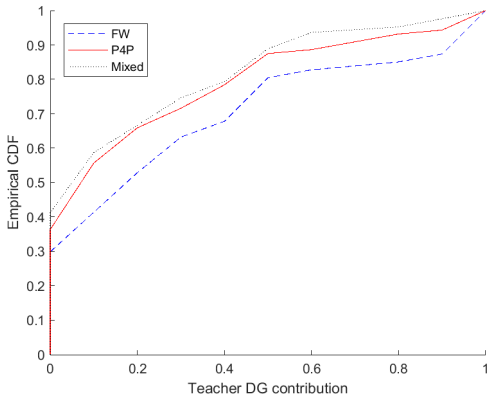
Our estimate of τ_A is -0.2069. The studentized coefficient has a standard deviation of 0.2083 under the sharp null. The randomization inference p-value is 0.31. It follows that we cannot reject the sharp null of no advertised P4P treatment effect on the observable skills of placed recruits, as measured by a skills test at baseline.

[22] Since this model assumes normality of the skill distribution, we fit item parameters using only the sample of incumbent teachers.

4.3 Hypothesis III. Advertised P4P induces differentially 'intrinsically' motivated recruits to be placed in schools.

Primary test In addition to the possibility that advertised P4P may select teachers on the basis of skill, such contracts may also change the distribution of intrinsic motivation among the pool of applicants. Our primary test uses teachers' allocation to others in the framed Dictator Game played at baseline, which we denote by x_{iqd} for teacher i with qualification q in district d. Figure 4 plots the distribution of this measure by advertised treatment arm.

Figure 4: Distribution of placed teacher Dictator Game contributions, by treatment arm



To test the sharp null of no effects we estimate a regression of the form

$$x_{iqd} = \tau_A T^A_{qd} + \gamma_q + \delta_d + e_{iqd}. \tag{5}$$

Our estimate of τ_A is -0.1079: teachers recruited under advertised P4P allocated approximately 10 percentage points of the stake less to the school on average in the framed Dictator Game. The studentized coefficient has a standard deviation of 0.0603 under the sharp null. The randomization inference p-value is 0.02. We can therefore reject this sharp null of no advertised P4P treatment effect on the intrinsic motivation of placed recruits at the 5 percent level.

4.4 Hypothesis IV. Advertised P4P induces the selection of higher (or lower) value-added teachers.

Measures of teacher skill and intrinsic motivation are policy relevant insofar as recruits with such favorable attributes are likely to deliver better learning outcomes for their students. To test whether this is the case, we combine experimental variation in the *advertised* contracts to which placed recruits applied, with the second-stage randomization in *experienced* contracts under which they worked. This allows us to estimate the impact of advertised P4P, holding constant the experienced contract: a pure compositional effect.

Primary test The measure of student learning which we deploy here, and in Section 4.5 below, is the empirical Bayes prediction of student ability, based on an IRT model of student assessments. This prediction is observed at the student-subject level, since each sampled student takes an assessment in all five core subjects. We denote this measure by z_{jbksr}, for student j in subject b, stream k, school s, and round r from which student j's outcome is drawn. (Since streams are nested within grades, we suppress an index for grades to reduce notation.) We standardize this measure of student learning within each grade, subject, and round so that it has a mean of zero and a standard deviation of one among students taught by incumbents in schools that experience the fixed-wage treatment.

The advertised treatment about which a given student's performance is informative depends on the identity of the teacher teaching that particular subject via qualification type and district. We denote this by T^A_{qd} for teacher i with qualification type q in district d, and suppress the dependence of the teacher's qualification q on the subject b, stream k, school s, and round r, which implies that $q = q(bksr)$. The experienced treatment is assigned at the school level, and is denoted by T^E_s.

Our primary test is for the impact of the advertised treatment on a recruits's annual value added, holding constant the actual (experienced) contractual treatment of the school into which they were placed. We pool data across the two years of intervention and estimate a specification of the type

$$z_{jbksr} = \tau_A T^A_{qd} + \tau_E T^E_s + \lambda_I I_i + \lambda_E T^E_s I_i + \rho_{br} \bar{z}_{ks,r-1} + \delta_d + \psi_r + e_{jbksr}, \tag{6}$$

for the learning outcome of student j in subject b, stream k, school s, and round r. We define $i = i(bksr)$ as an identifier for the teacher assigned to that subject-stream-school-round. The variable I_i is an indicator for whether the teacher is an incumbent, and the index $q = q(i)$ denotes the qualification type of teacher i if that teacher is a recruit (and is undefined if the teacher is an incumbent, so that T^A_{qd} is always zero for incumbents). The variable $\bar{z}_{ks,r-1}$ denotes the vector of average outcomes in the once-lagged assessment among students now placed in that stream, and its coefficient, ρ_{br} is subject- and round-specific. As determined in our pre-analysis plan, we estimate this model by a linear mixed effects model, allowing for normally distributed random effects at the student-round level. This specification seeks to maximize power for the ability to reject the null of no advertised treatment effects by pooling recruits placed in experienced P4P and experienced FW treatments. Note that the specification in (6) is relevant to both Hypothesis IV via τ_A and Hypothesis V via τ_E. We discuss only τ_A here, postponing discussion of τ_E until the next section.

As the first column of Table 7 reports, our estimate of τ_A is close to zero at 0.01 standard deviations of pupil learning. The studentized coefficient has a standard deviation of 0.0263 under the sharp null. Comparing this studentized estimate with its randomization inference distribution yields a p-value of 0.56, indicating that we cannot reject the null of no impact of advertised P4P on teacher value added. However, since randomization inference is well-powered, we do feel confident in ruling out *negative* selection effects.

Secondary tests We begin our secondary tests by allowing for heterogeneous treatment impacts by round. Figure 5 illustrates the evolution of the advertised, experienced and combined effect of P4P contracts over time. As we discuss below, the combined effect increases from Year 1 to 2, primarily due to dynamics in the experienced P4P treatment effect. But by Year 2 recruits brought in under advertised P4P also begin to outperform those brought in under advertised FW contracts. As the second column in Table 7 reports, the estimate of τ_A for Year 1 (equivalently round 1) is -0.03σ with a randomization inference p-value of 0.21. In Year 2, this estimate is 0.05σ with a randomization inference p-value of 0.12.

The theoretical framework set out in the pre-analysis plan makes clear that the impact of advertised treatment on teacher value-added should depend on the contractual environment into which recruits are placed. Consequently, we also estimate a secondary specification that allows

Figure 5: Impacts on pupil learning, by round

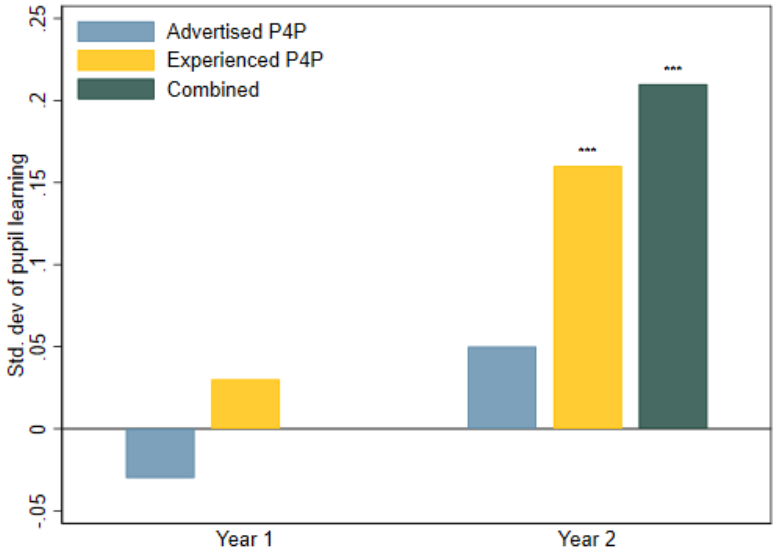


Table 7: Impacts on pupil learning (teacher value-added)

	Pooled	Round 1	Round 2	Interacted
τ_A	0.01	-0.03	0.05	0.01
	[0.56]	[0.21]	[0.12]	[0.60]
τ_E	0.09	0.03	0.16	0.11
	[0.01]	[0.36]	[0.00]	[0.00]
τ_{AE}				-0.01
				[0.81]
λ_E	-0.06	-0.02	-0.10	-0.07
	[0.04]	[0.56]	[0.01]	[0.03]
$\tau_A + \tau_{AE}$				0.00
				[0.87]
$\tau_E + \tau_{AE}$				0.09
				[0.05]
$\tau_E + \lambda_E$	0.04	0.01	0.06	0.04
	[0.14]	[0.71]	[0.07]	[0.14]

Notes: For each estimated parameter, or combination of parameters, the table reports the point estimate, stated in standard deviations of pupil learning, together with its randomization inference p-value. Randomization inference is conducted on the associated z statistic, with 2,000 permutations of the relevant treatment assignment.

advertised treatment effects on teacher value-added to differ by experienced treatment, including an interaction term between the two treatments. This interacted model takes the form

$$z_{jbkgsr} = \tau_A T^A_{qd} + \tau_E T^E_s + \tau_{AE} T^A_{qd} T^E_s + \lambda_I I_i + \lambda_E T^E_s I_i + \rho_{bgr} \bar{z}_{ks,r-1} + \delta_d + \psi_r + e_{jbksr}. \quad (7)$$

Here, the compositional effect of advertised P4P among recruits placed in FW schools is given by τ_A (a comparison of on-the-job performance across groups a and b, as defined in Figure 1). Likewise, the compositional effect of advertised P4P among recruits placed in P4P schools is given by $\tau_A + \tau_{AE}$ (a comparison of groups c and d). As the fourth column of Table 7 reoprts, in this interacted model the impact of advertised P4P is actually smaller—negative in point-estimate terms—when these recruits are placed in P4P schools, relative to when they experience a FW contract. However, undertaking a randomization inference procedure that permutes *both* treatment assignments $\mathcal{T}^{\mathcal{A}}$ and $\mathcal{T}^{\mathcal{E}}$, we cannot reject the sharp null that these responses are the same, i.e. $\tau_{AE} = 0$.

4.5 Hypothesis V. Experienced P4P creates incentives which contribute to higher (or lower) teacher value-added.

Primary test Here, we test for an impact of experienced P4P on recruits' annual value-added, holding constant the advertised treatment under which they applied. We pool data across the two years and use the specification in (6). As the first column of Table 7 reports, our estimate of τ_E is 0.09σ. The studentized coefficient has a standard deviation of 0.06 under the sharp null. Comparing this studentized estimate with its randomization inference distribution (here permuting only $\mathcal{T}^{\mathcal{E}}$) yields a p-value of 0.01, implying that we can reject the sharp null of no experienced P4P treatment effect on placed recruits at the 1 percent level.

Secondary tests Figure 5 shows that there is only a small impact of experienced P4P in Year 1 but a much more sizeable impact by Year 2. As reported in the second column of Table 7, the estimate of τ_E for Year 1 is 0.03σ with a randomization inference p-value of 0.36. In Year 2, this coefficient is over five times larger at 0.16σ with a randomization inference p-value of 0.00. This second year impact corresponds to moving a student from the median (50th percentile) up to the 56th percentile of the student learning distribution—a modest but certainly economically

meaningful result.

Table 7 also reports coefficients from the interacted specification in (7). In this model, τ_E gives the incentive effect of experienced P4P among recruits who applied under FW contractual conditions (a comparison of groups a and c, as defined in Figure 1), while $\tau_E + \tau_{AE}$ gives the incentive effect of experienced P4P among recruits who applied under P4P contractual conditions (a comparison of groups b and d). As the fourth column of Table 7 shows, the point estimate of τ_E is slightly larger than that of $\tau_E + \tau_{AE}$ but we cannot reject the sharp null that these responses are the same. In Section 2.2, we noted that the re-randomization could in principle have caused 'disappointment effects'. If this is true then, because it is groups b and c who received the 'surprise', τ_E should actually be *smaller* than $\tau_E + \tau_{AE}$. Since if anything we find the reverse, we feel that it is unlikely that the re-randomization itself had a causal impact on behaviour.

4.6 Hypothesis VI. Selection and incentive effects are apparent in the 4P performance metric.

Sections 4.4 and 4.5 speak to the obvious policy question, namely whether there are impacts of advertised and experienced P4P contracts on a direct measure of student learning. This measure was not included in the P4P contract, however. For completeness, and to gain an understanding into mechanisms, in this subsection we study whether there are impacts on the *contracted* metrics, i.e. on the 4Ps.

Primary test For our primary test, we focus on the composite P4P teacher performance metric. We pool data across the two years and use a specification of the type

$$m_{iqsdr} = \tau_A T^A_{qd} + \tau_E T^E_s + \lambda_I I_i + \lambda_E T^E_s I_i + \gamma_q + \delta_d + \psi_r + e_{iqsdr}, \tag{8}$$

for the composite performance metric of teacher i with qualification q in school s of district d, as observed in post-treatment round r. As above, the variable I_i is an indicator for whether the teacher is an incumbent (recall that T^A_{qd} is always zero for incumbents).[23] This pooled specification provides our most powerful available test for violations of the sharp null hypothesis of no advertised

[23]Note that any attribute of recruits themselves, even if observed at baseline, suffers from the 'bad controls' problem. Observed values of this covariate could be an outcome of the advertised treatment.

P4P effect ($\tau_A = 0$), if the truth is that there are no differences in impact across experienced treatment arms. Likewise, it provides our most powerful available test for violations of the sharp null hypothesis of no experienced P4P effect ($\tau_E = 0$), if the truth is that there are no differences in impact across advertised treatment arms. Note that a linear mixed effects model with student-level random effects is no longer applicable: outcomes are constructed at the teacher-level, and given their rank-based construction, normality does not seem a helpful approximation to the distribution of error terms. As determined in our pre-analysis plan, we therefore estimate equation (8) with a round-school random-effects estimator to improve efficiency. The permutations of treatments used for inferential purposes mirror those in Hypothesis IV and V.

Results from estimating equation (8) are reported in the first column of Table 8. Our estimate of τ_A is -0.04 percentile ranks. Comparing this studentized estimate with its randomization inference distribution (here permuting only $\mathcal{T}^A$) yields a $p-$value of 0.11. The confidence interval, based on inversion of the RI test, is $[-0.10, 0.02]$. Our estimate of τ_E is substantially larger at 0.24 percentile ranks. Comparing this studentized estimate with its randomization inference distribution (here permuting only $\mathcal{T}^E$) yields a $p-$value of 0.00. The confidence interval, based on inversion of the RI test, is $[0.19, 0.28]$. Hence, while we cannot reject the null of no impact of advertised P4P, we can reject the null of no impact of experienced P4P on the composite 4P metric.

Secondary tests Our secondary tests repeat the analysis for: each sub-component of the composite 4P performance metric; for Year 2 data, and for interacted specifications. The second through fifth columns in Panel A of Table 8 report results from specifications where we replace the composite measure m_{iqsdr} from equation (8) with the following outcomes: the fraction of spot-check days in post-treatment round r on which teacher i, with qualification q, in school s of district d is observed to be present at the start of the school day (column 2); a binary indicator of whether teacher i, with qualification q, in school s of district d has a lesson plan on a randomly chosen spot-check day in post-treatment round r (column 3); the classroom observation score, measured on a four-point scale, of teacher i, with qualification q, in school s of district d in post-treatment round r (Column 4); and the Barlevy-Neal pupil learning percentile rank (BN rank) of teacher i, with qualification q, in school s of district d in post-treatment round r. Panel B of Table 8 reports results using Year

2 data only. Panel C returns to the pooled data but uses the following interacted specification

$$m_{iqsdr} = \tau_A T^A_{qd} + \tau_E T^E_s + \tau_{AE} T^A_{qd} T^E_s + \lambda_I I_i + \lambda_E T^E_s I_i + \gamma_q + \delta_d + \psi_r + e_{iqsdr} \quad (9)$$

first for the composite performance metric, and then replacing m_{iqsdr} with each of the four sub-components.

Echoing the findings in Section 4.4, our estimates of τ_A increase moving from the pooled to Year 2 specification but are never statistically or economically significantly different from zero. This is true for the composite metric and for each of the four sub-components. Since randomization inference remains well-powered, we interpret this as evidence that advertised P4P did not have a *negative* effect on teachers' inputs into pupil learning, or on teachers' outputs as captured by BN rank. Likewise, the interaction effects between treatments are never statistically significant, indicating that there is no evidence that the impact of advertised P4P on these metrics differed between recruits placed into P4P versus FW schools.

Echoing the findings in Section 4.5, our estimates of τ_E also increase moving from the pooled to Year 2 specification. For both the composite metric and three of the four sub-components these estimates are always positive and statistically significant at conventional levels. The exception is teacher preparation, where (quite possibly due to the noisy nature of this measure) there is no evidence of an effect, either in the pooled or Year 2 specification. Again, the interaction effects between treatments are never statistically significant.

We take the following points away from this secondary analysis. First, the specifications using the teacher-level BN rank as the dependant variable confirm our results from earlier specifications using the direct, student-level measure of learning. Second, the specifications using teacher inputs as the dependant variable suggest that these impacts on pupil learning may be driven (at least in part) by improvements in teacher presence and pedagogy. In Year 2, teacher presence was 6 percentage points higher among recruits who experienced the P4P contract compared to recruits who experienced the FW contract. (A sizeable impact given that baseline teacher presence was already 90 percent.) Moreover, in the same year, recruits who experienced P4P were more effective in their classroom practices than recruits who received a fixed-wage by 0.26 points, as measured on the four-point scale.

Table 8: Estimated effects on dimensions of the composite 4P performance metric

	Preparation	Presence	Pedagogy	Pupil learning	Summary metric
Model A: Direct effects only					
Advertised P4P	0.07	0.00	0.02	-0.02	-0.04
	[-0.19, 0.38]	[-0.07, 0.08]	[-0.09, 0.11]	[-0.10, 0.03]	[-0.10, 0.02]
	(0.36)	(0.95)	(0.42)	(0.28)	(0.11)
Experienced P4P	0.24	0.02	0.28	0.09	0.24
	[-0.15, 0.18]	[0.01, 0.15]	[-0.02, 0.22]	[0.02, 0.16]	[0.19, 0.28]
	(0.80)	(0.01)	(0.06)	(0.00)	(0.00)
Experienced P4P × Incumbent	0.01	-0.00	0.03	-0.00	0.03
	[-0.04, 0.19]	[-0.08, 0.05]	[-0.02, 0.18]	[-0.05, 0.04]	[-0.01, 0.07]
	(0.19)	(0.74)	(0.10)	(0.92)	(0.09)
Model B: Interactions between advertised and experienced contracts					
Advertised P4P	0.16	-0.01	0.12	-0.00	-0.03
	[-0.21, 0.54]	[-0.19, 0.20]	[-0.39, 0.65]	[-0.15, 0.13]	[-0.13, 0.07]
	(0.22)	(0.88)	(0.44)	(0.92)	(0.43)
Experienced P4P	0.24	0.02	0.36	0.08	0.23
	[-0.29, 0.29]	[-0.02, 0.18]	[-0.07, 0.40]	[-0.00, 0.17]	[0.14, 0.30]
	(0.96)	(0.07)	(0.12)	(0.04)	(0.00)
Advertised P4P × Experienced P4P	-0.11	0.02	-0.12	-0.03	-0.02
	[-0.49, 0.29]	[-0.14, 0.17]	[-0.51, 0.31]	[-0.17, 0.10]	[-0.12, 0.09]
	(0.57)	(0.73)	(0.54)	(0.62)	(0.73)
Experienced P4P × Incumbent	0.01	0.00	-0.05	0.00	0.04
	[-0.09, 0.29]	[-0.11, 0.08]	[-0.15, 0.17]	[-0.06, 0.07]	[-0.01, 0.11]
	(0.30)	(0.82)	(0.95)	(0.90)	(0.08)
Observations	2512	3453	2134	3048	3995
FW recruit mean (SD)	0.64	0.89	1.98	0.48	0.49
	(0.49)	(0.32)	(0.57)	(0.27)	(0.22)
FW incumbent mean (SD)	0.50	0.87	2.05	0.45	0.37
	(0.50)	(0.33)	(0.49)	(0.28)	(0.24)

Notes: For each estimated parameter, the table reports the point estimate; 95 percent confidence interval in brackets, and p-value in parentheses. Randomization inference, conducted on the associated t statistic, is undertaken with 2,000 alternative permutations of treatment. All estimates pool years one and two, but outcomes are observed in the FW arm during Year 2 only.

5 Exploratory results on dynamics

Our two-tier experiment was designed to evaluate the impact of pay-for-performance and, in particular, to quantify the relative importance of a compositional margin at the recruitment stage versus an effort margin on-the-job. The hypotheses specified in our pre-analysis plan (Table 6) refer to selection-in and incentives among placed recruits. Since within-year teacher turnover was limited by design and within-year changes in teacher skill and motivation are likely small, the total effect of P4P in Year 1 can plausibly only be driven by a change in the type of teachers recruited and/or a change in effort resulting from the provision of extrinsic incentives.

Interpreting the total effect of P4P in Year 2 is more complex, however. First, we made no attempt to discourage *between*-year teacher turnover, and so there is the possibility of a further compositional margin at the retention stage (c.f. Muralidharan and Sundararaman 2011). Experienced P4P may have selected-out the low skilled (Lazear, 2000) or, more pessimistically, the highly intrinsically motivated. Second, given the longer time frame, teacher characteristics could have changed. Experienced P4P may have eroded a given teacher's intrinsic motivation (as hypothesised in the largely theoretical literature on motivational crowding out) or, more optimistically, encouraged a given teacher to improve her classroom skills. In this section, we conduct an exploratory analysis of these dynamic effects.[24]

5.1 Retention effects

Does experienced P4P affect retention rates among recruits? To answer this question we look for impacts on the likelihood that a recruit is still employed at midline in February 2017 at the start of the Year 2 (i.e. after experiencing P4P in Year 1, although before the performance awards were announced). Our primary test of this hypothesis is a linear probability model of the form

$$\Pr[employed_{iqd2} = 1] = \tau_E T_s^E + \gamma_q + \delta_d, \tag{10}$$

where $employed_{iqd2}$ is an indicator variable taking a value of one if teacher i with subject-family qualification q in district d is still employed by the school at the start of Year 2, and γ_q and

[24] We emphasise that this material in exploratory; the hypotheses tested in this section were not part of our pre-analysis plan. That said, the structure of the analysis in this section does follow a related pre-analysis plan (intended for a companion paper) which we uploaded to our trial registry on October 3 2018 *prior* to unblinding of our data.

δ_d are the usual subject-family qualification and district indicators. As Column 1 of Table 9 reports, our estimate of τ_E is zero with a randomization inference p-value of 0.96. There is no statistically significant impact of experienced P4P on the retention rate of recruits; the retention rate is practically identical—at around 80 percent—among recruits experiencing P4P and those experiencing FW.

Table 9: Retention of new recruits

	(1)	(2)	(3)
Experienced P4P	0.00	-0.04	-0.08
	[0.96]	[0.41]	[0.23]
Interaction		-0.05	0.16
		[0.38]	[0.36]
Heterogeneity by...		Test score	DG share sent
Observations	249	238	238

Notes: Randomization inference $p-$values in brackets, representing 2,000 draws of the experienced treatment. All specifications include controls for districts and subjects of teacher qualification.

It is worth noting that there is also no impact of experienced P4P on *intentions* to leave in Year 3. In the endline survey in November 2017, we asked teachers the question:"How likely is it that you will leave your job at this school over the coming year?". Answers were given on a 5-point scale. For analytical purposes we collapse these answers into a binary indicator coded to 1 for 'very likely' or 'likely' and 0 otherwise, and estimate specifications analogous to equations (8) and (9). As the first column of Table B.1 shows, there is no statistically significant impact of experienced P4P on recruits' self-reported likelihood of leaving in Year 3. Our estimate of τ_E is -0.06 with a randomization inference $p-$value of 0.40.[25]

Of course, 20 percent attrition from Year 1 to Year 2 is non-negligible. And the fact that retention *rates* are similar does not rule out the possibility of an impact of experienced P4P on the *type* of recruits retained. We turn to this issue below.

Does experienced P4P induce differentially skilled recruits to be retained? To answer this question we use a teacher's performance on the baseline Grading Task in the primary subject

[25]It is notable that RI is less well-powered than in our main specifications, perhaps indicating the noisy nature of self-reported measure.

he/she teaches (see Section 3) to obtain an IRT estimate of his/her ability in this subject, denoted z_i, and estimate an interacted model of the form

$$\Pr[employed_{iqd2} = 1] = \tau_E T_s^E + \zeta T_s^E z_i + \beta z_i + \gamma_q + \delta_d. \tag{11}$$

Inference for the key parameter, ζ, is undertaken by performing randomization inference for alternative assignments of the school-level experienced treatment indicator. As Column 2 of Table 9 reports, our estimate of ζ is -0.05, with a randomization inference p-value of 0.38. There is not a significant difference in selection-out on baseline teacher skill across the experienced treatments. In other words, there is no evidence that experienced P4P induces differentially skilled recruits to be retained. To aid intepretation, Figure A.2 shows two plots of fitted values from a simple (bivariate) linear regression of $employed_{iqd2}$ on z_i. The red, dashed plot uses data on recruits in the experienced P4P arm, while the blue dotted plot is for recruits in the experienced FW arm. Both plots are upward sloping indicating that it is the low skilled teachers who select-out between Year 1 and 2. Interestingly, this positive relationship is stronger, and indeed is only statistically significantly different from zero, among recruits who experienced FW.

Does experienced P4P induce differentially intrinsically motivated recruits to be retained? To answer this question we use the contribution sent in the framed Dictator Game played by all recruits at baseline, denoted x_i, we re-estimate the interacted model in equation (11), replacing z_i with x_i. As Column 3 of Table 9 reports, our estimate of ζ in this specification is 0.16, with a randomization inference p-value of 0.36. There is not a significant difference in selection-out on baseline teacher intrinsic motivation across the experienced treatments. In other words, there is also no evidence that experienced P4P induces differentially intrinsically motivated recruits to be retained. To aid interpretation, Figure A.3 shows two plots of fitted values from a simple (bivariate) linear regression of $employed_{iqd2}$ on x_i. The red, dashed plot uses data on recruits in the experienced P4P arm, while the blue dotted plot is for recruits in the experienced FW arm. Both plots are downward sloping indicating that it is the high motivation teachers who select-out between Year 1 and 2. Again, the effect is stronger among recruits who experienced FW.

Table 10: Recruit characteristics at endline

	Teacher score	DG send
Experienced P4P	0.57	-0.04
	[0.63]	[0.06]
Observations	170	169

Notes: Randomization inference $p-$values in brackets, representing 2,000 draws of the experienced treatment. All specifications include controls for district and subject-of-qualification.

5.2 Changes in retained teacher characteristics

To assess whether experienced P4P changes within-retained-recruit teacher skill or intrinsic motivation from baseline to endline, we estimate the following ANCOVA specification

$$y_{isd2} = \tau_E T_s^E + \rho y_{isd0} + \gamma_q + \delta_d + e_{isd}, \tag{12}$$

where y_{iqsd2} is the characteristic (raw Grading Task score or framed Dictator Game contribution) of retained recruit i with qualification q in school s and district d at endline (round 2), and y_{iqsd0} is this characteristic of retained recruit i at baseline (round 0).[26] As Column (1) of Table 10 reports, our estimate of τ_E in the endline Grading Task score specification is 0.57, with a randomization inference $p-$value of 0.63. Our estimate of τ_E in the framed Dictator Game contribution specification is -0.04, with a randomization inference $p-$value of 0.06. Both estimates are small in magnitude and reject the sharp null only at the 10 percent level, in the case of the Dictator Game allocations to others. Hence, to the extent that contributions in the Dictator Game are positively associated with teachers' intrinsic motivation, we find no evidence that the *rising* effects of experienced P4P from Year 1 to Year 2 are driven by *positive* changes in within-retained-recruit teacher skill or intrinsic motivation, at least on these metrics.

Before moving on, it is worth noting that this Dictator Game result could be interpreted as weak evidence that the experience of P4P contracts crowded out the intrinsic motivation of recruits. Sadly, we do not have any related measures taken at both baseline and endline with which to further probe *changes* in motivation. However, we do have a range of related measures, specifically

[26] The raw Grading Task score is measured on a scale of 0-30. To be replaced by IRT estimate in due course.

indexes of public sector motivation, job satisfaction, and positive/negative affect taken in Year 2.[27] As the second through fifth columns of Table B.1 show, there is no statistically significant impact of experienced P4P on any index. Although the randomization inference confidence intervals are not tight (presumably reflecting the noisy nature of these survey measures), we can rule out economically meaningful negative impacts of experienced P4P on these measures of motivation.

5.3 Decomposing the total effect of P4P in Year 2

The *total effect* of the P4P contract in Year 2 combines both the advertised and experienced impacts: $\tau_A + \tau_E$. Our estimate is $0.05\sigma + 0.16\sigma = 0.21\sigma$ which is statistically significant at the 1 percent level, as indicated in Figure 5. Given our two-tier experimental design, we can say unequivocally that nearly a quarter of the total effect of P4P in Year 2 is accounted for by selection-in at the recruitment stage—a compositional margin that is not typically identified by studies of performance-pay. The remainder of the total effect could be due to selection-out at the retention stage, changes in teacher skill within-retained recruits, changes in effort, or all three. We cannot exploit a design feature to disentangle these effects and so instead turn to the results in Section 5.1 and 5.2. On this basis, it seems unlikely that our estimate of τ_E is driven by selection-out at the retention stage—the compositional margin famously highlighted by Lazear (2000). A teacher's baseline Grading Task score is arguably a more plausible driver of pupil learning than his/her baseline intrinsic motivation as measured by the framed Dictator Game contribution. And we find statistically significant selection-out of low baseline skill teachers in the experienced FW arm but not in the experienced P4P arm. If anything, *selection-out runs the wrong way* to explain τ_E. In view of Table 10, it also seems unlikely that τ_E is driven by a change in teacher skill from baseline to endline within-retained recruits. We therefore attribute the remaining three quarters of the total effect of the P4P contract in Year 2 to increased effort on-the-job. Of course, this effort margin is net of any motivational crowding out from P4P. To the extent that contributions in the Dictator Game are positively associated with teachers intrinsic motivation, Table 10 suggests that this potential countervailing effect is small in our context.

[27] We follow Dal Bó et al. (2013) in using the Perry Index of public service motivation. Our measure was constructed using data from telephone interviews in Summer 2017. We follow Bloom et al. (2015) in using the Maslach Burnout Index as a way to capture job satisfaction and the Clark-Tellgen Index of positive and negative affect as a way to capture the overall attitude of teachers. These measures were constructed using data from the endline teacher survey.

6 Conclusion

This study has reported on the results of a randomized, controlled trial designed to test for both the compositional and effort-margin responses of a pay-for-performance contract in Rwandan primary schools. The study's unique two-tier design allows us to decompose the total effect of performance pay into these constituent parts.

Drawing on a theoretical model we defined a narrow set of hypotheses to test in our pre-analysis plan, and used simulations on blinded data to develop high-powered tests of these hypotheses before carrying out the main analysis.

In terms of *recruitment*, we find that advertisement of the P4P contract changed the composition of the teaching workforce, drawing in individuals who were more money-oriented. Specifically, teachers recruited under P4P exhibit less other-regarding preferences in a framed Dictator Game designed to measure intrinsic motivation. However, these recruits were not less effective teachers. Pooled estimates of the impact of advertised P4P on teacher value-added are precisely estimated, and close to zero. If anything, in the second year of the study—when differences in teaching efficacy among recruits emerge more clearly—there is a positive selection effect. Turning to the *effort* margin, we find significant and economically large effects of experience P4P on learning outcomes. This learning gain is apparent both in traditionally constructed value added, as well as in the Barlevy-Neal percentile rank metric that was used in the P4P contract. It is further reflected by gains in both teacher presence and pedagogy, both of which were part of the composite '4P' performance metric. We find no statistically significant differences in teacher attrition across study schools, and detect no negative selection effects of P4P on this *retention* margin.

In sum, we find that the recruitment, effort, and retention-margins of performance contracts combine to raise learning quality. In the second year of the study, we estimate the total effect of P4P to be 0.21 standard deviations of pupil learning. One quarter of this impact can be attributed to selection at the recruitment stage, with the remaining three-quarters arising from increased effort.

The above results provide proof of concept in an externally implemented and funded research environment. There are several potential problems when contemplating a move to a government-led, nationwide scheme. The first question is whether the necessary *measurement* could be conducted by the government on a national scale. In terms of pupil learning, the minimum requirement is a system

of repeated annual assessments across grades and key subjects. Such a system does not yet exist in Rwanda but will soon be introduced, as part of the recently announced 'comprehensive assessment' program.[28] Measurement of the other 'Ps'— teacher presence, preparation, and pedagogy—is less complex and could in principle be conducted by head teachers (or other existing school or district staff members) at limited cost.

A second question is whether there is scope in the government budget to cover the *incentive payments*. The FW contract (RWF 20,000 for all teachers) was equivalent to a 3 percent increase in an average teacher's salary. This 3 percent figure is broadly in line with annual increments in other sectors under the *imihigo* system, as well a recently floated (fixed) wage increase for teachers. Since the P4P contract was designed to be equivalent in costs for salary, this suggests that the incentive payments could be introduced within the current budget envelope.

A third question is whether there will be support for P4P from key stakeholders. Opponents of P4P often argue that it is unpopular with teachers. To investigate this issue, we included a question on teacher attitudes to P4P in our endline survey. Specifically, we asked teachers: "What is your overall opinion about the idea of providing high-performing teachers with bonus payments on the basis of objective measures of student performance improvement?"[29] Table 11 reports the distribution of teacher responses. Among the group of recruits who applied under the advertised FW treatment, 82.8 percent held a positive view of P4P (answers in the top two categories). Notably, the proportion holding a positive view was higher in the experienced P4P subgroup than in the experienced FW subgroup (90.4 vs. 75.7 percent). We see a similar pattern among incumbents who were in post prior to the advertised treatment: 80.3 percent held a positive view of P4P, and the proportion holding a positive view was higher in the experienced P4P subgroup than in the experienced FW subgroup (82.1 vs. 78.2 percent). Among the group of recruits who applied under the advertised P4P treatment, 83.4 percent held either a positive view of P4P. Here, the proportion holding a positive view was slightly lower, although not statistically significantly so, in

[28]On January 28, 2019 the Government of Rwanda announced a Cabinet resolution establishing comprehensive assessment for all basic level education in Rwanda. In a letter dated June 18, 2019 Education Minister Dr. Eugene Mutimura issued new guidelines on how to conduct student assessment in primary, secondary, technical and vocational training schools as well as teacher training colleges. See https://www.newtimes.co.rw/opinions/mineducs-new-guide-student-assessment-triggers-debate.

[29]Here, we follow the phrasing used in the teacher surveys conducted by Muralidharan and Sundararaman (2011a). Answers were possible on a 5-point Likert scale, ranging from 'very unfavourable' to 'very favourable'.

Table 11: Teacher attitudes toward P4P at endline

	Very unfavorable	Somewhat unfavorable	Neutral	Somewhat favourable	Very favourable
Recruits applying under FW (64)	4.7%	4.7%	7.8%	10.9%	71.9%
—Experiencing FW (33)	6.1%	9.1%	9.1%	3.0%	72.7%
—Experiencing P4P (31)	3.2%	0.0%	6.5%	19.4%	71.0%
Recruits applying under P4P (60)	5.0%	3.3%	8.3%	1.7%	81.7%
—Experiencing FW (32)	6.3%	0.0%	6.3%	0.0%	87.5%
—Experiencing P4P (28)	3.6%	7.1%	10.7%	3.6%	75.0%
Incumbent teachers (1,113)	5.0%	7.5%	7.2%	9.9%	70.4%
—Experiencing FW (537)	5.2%	8.6%	8.0%	8.6%	69.6%
—Experiencing P4P (576)	4.9%	6.6%	6.4%	11.1%	71.0%

Notes: The table reports the distribution of answers to the following question on the endline teacher survey: "What is your overall opinion about the idea of providing high-performing teachers with bonus payments on the basis of objective measures of student performance improvement?" Figures in parentheses give the number of respondents in each treatment category.

the experienced P4P subgroup than in the experienced FW subgroup (78.6 vs. 87.5 percent).[30] We conclude from these results that the concept of P4P *is* popular with teachers, both in principle and in practice after having experienced it.

Drawing this discussion together, it seems possible that potential measurement, budget, and political challenges can be overcome and that prospects for scale-up in Rwanda are good.

[30]Note that there is no evidence that the re-randomization resulted in hostility toward P4P—if anything the reverse.

Acknowledgements

We thank counterparts at REB and MINEDUC for their advice and collaboration. We are grateful to Katherine Casey, Erika Deserranno, David Evans, Dean Eckles, Frederico Finan, Macartan Humphreys, Pam Jakiela, Julien Labonne, David McKenzie, Berk Özler, and Cyrus Samii for helpful conversations and comments. This project would not have been possible without the contributions of numerous IPA staff members, including Kris Cox, Stephanie De Mel, Olive Karekezi Kemirembe, Doug Kirke-Smith, Emmanuel Musafiri, and Phillip Okull. Claire Cullen, Robbie Dean, Ali Hamza, Gerald Ipapa, and Saahil Karpe provided excellent research assistance. Financial support for the research on this project was provided by the U.K. Department for International Development (DfID) via the International Growth Centre and the Economic Development and Institutions Programme, Oxford University's John Fell Fund, and the World Bank's Strategic Impact Evaluation Fund (SIEF) and REACH trust fund. The findings in this report are the opinions of the authors, and do not represent the opinions of the World Bank, its Executive Directors, or the governments they represent. All errors and omissions are our own.

References

Adnot, Melinda, Thomas Dee, Veronica Katz, and James Wyckoff, "Teacher turnover, teacher quality, and student achievement in DCPS," *Educational Evaluation and Policy Analysis*, 2017, *39* (1), 54–76.

Almlund, Mathilde, Angela Lee Duckworth, James Heckman, and Tim Kautz, "Personality psychology and economics," in E. A. Hanushek, S. Machin, and L. Woessmann, eds., *Handbook of the Economics of Education*, Vol. 4, Amsterdam: Elsevier, 2011, pp. 1–181.

Anderson, Michael L and Jeremy Magruder, "Split-sample strategies for avoiding false discoveries," NBER Working Paper No. 23544 6 2017.

Ashraf, Nava, James Berry, and Jesse M Shapiro, "Can higher prices stimulate product use? Evidence from a field experiment in Zambia," *American Economic Review*, December 2010, *100* (5), 2382–2413.

_ , Oriana Bandiera, and B.Kelsey Jack, "No margin, no mission? A field experiment on incentives for public service delivery," *Journal of Public Economics*, December 2014, (120), 1–17.

_ , _ , and Scott S Lee, "Do-gooders and go-getters: Selection and performance in public service delivery," Working paper June 2016.

Barlevy, Gadi and Derek Neal, "Pay for percentile," *American Economic Review*, August 2012, *102* (5), 1805–1831.

Bénabou, Roland and Jean Tirole, "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, 2003, *70*, 489–520.

Biasi, Barbara, "The labor market for teachers under different pay schemes," NBER Working Paper No. 24813 3 2019.

Binswanger, Hans P, "Attitudes toward risk: Experimental measurement in rural India," *American Journal of Agricultural Economics*, August 1980, *62* (3), 395–407.

Bloom, Nicholas, James Liang, John Roberts, and Zhichun Jenny Ying, "Does working from home work? Evidence from a Chinese experiment," *Quartely Journal of Economics*, 2015, pp. 165–218.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane, "Enrollment without learning: Teacher effeort, knowledge, and skill in primary schools in Africa," *Journal of Economic Perspectives*, Summer 2017, *31* (4), 185–204.

Brock, Michelle, Andreas Lange, and Kenneth Leonard, "Generosity and Prosocial Behavior in Healthcare Provision: Evidence from the Laboratory and Field," *Journal of Human Resources*, April 2016, *51* (1), 133–162.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek, "Gender, competitiveness, and career chocies," *Quarterly Journal of Economics*, 8 2014, *129* (3), 1409–1447.

Callen, Michael, Saad Gulzar, Ali Hasanain, Yasir Khan, and Arman Rezaee, "Personalities and public sector performance: Evidence from a health experiment in Pakistan," NBER Working Paper No. 21180 4 2018.

Chetan, Dave, Catherine C. Eckel, Cathleen A. Johnson, and Christian Rojas, "Eliciting risk preferences: When is simple better?," *Journal of Risk and Uncertainty*, November 2010, *41*, 219–243.

Chetty, Raj, John N Friedman, and Jonah E Rockoff, "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates," *American Economic Review*, 2014.

_ , _ , and _ , "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood," *American Economic Review*, September 2014, *104* (9), 2633–2679.

Cohen, Jessica and Pascaline Dupas, "Free distribution or cost-sharing? Evidence from a Randomized Malaria Prevention Experiment," *Quarterly Journal of Economics*, February 2010, *125* (1), 1–45.

Dal Bó, Ernesto and Frederico Finan, "At the Intersection: A Review of Institutions in Economic Development," *EDI Working Paper*, 2016.

_ , _ , and Martin Rossi, "Strengthening state capabilities: The role of financial incentives in the call to public service," *Quarterly Journal of Economics*, 2013, *128* (3), 1169–1218.

Danielson, Charlotte, *Enhancing professional practice: A framework for teaching*, 2 ed., Alexandria, VA: Association for Supervision and Curriculum Development, 2007.

Dee, Thomas and James Wyckoff, "Incentives, selection, and teacher performance: Evidence from IMPACT," *Journal of Policy*, 2015, *34* (2), 267–297.

Delfgaauw, Josse and Robert Dur, "Incentives and Workers' Motivation in the Public Sector," *Economic Journal*, January 2007, *118* (525), 171–191.

Deserranno, Erika, "Financial incentives as signals: experimental evidence from the recruitment of village promoters in Uganda," *American Economic Journal: Applied Economics*, 2019, *11* (1), 277–317.

Donato, Katherine, Grannt Miller, Manoj Mohanan, Yulya Truskinovsky, and Marcos Vera-Hernández, "Personality traits and performance contracts: Evidence from a field experiment among maternity care providers in India," *American Economic Review*, 2017, *107* (5), 506–510.

Eckel, Catherine and Philip Grossman, "Altruism in anonymous dictator games," *Games and Economic Behavior*, September 1996, *16* (1), 181–191.

_ **and** _ , "Forecasting risk attitudes: An experimental study using actual and forecast gamble choices," *Journal of Economic B*, 2008, *68* (1), 1–17.

EunYi Chung and Joseph P Romano, "Exact and asymptotically robust permutation tests," *The Annals of Statistics*, 2013, *41* (2), 488–507.

Fafchamps, Marcel and Julien Labonne, "Using split samples to improve inference on causal effects," *Political Analysis*, 2017, *25*, 465–482.

Gensowski, Miriam, "Personality, IQ, and lifetime earnings," *Labour Economics*, 2017, *51*, 170–183.

Gilligan, Dan, Naureen Karachiwalla, Ibrahim Kasirye, Adrienne Lucas, and Derek Neal, "Educator Incentives and Educational Triage in Rural Primary Schools," NBER Working Paper No. 24911 August 2018.

Glewwe, Paul, Nauman Ilias, and Michael Kremer, "Teacher incentives," *American Economic Journal: Applied Economics*, July 2010, *2* (3), 205–227.

Hanushek, Eric A and Ludger Woessmann, "Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation," *Journal of Economic Growth*, 2012, *17*, 267–321.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt, "Fishing, commitment, and communication: A proposal for comprehensive nonbinding research Registration," *Political Analysis*, 2013, *21* (1), 1–20.

John, Oliver P, "The 'Big Give' factor taxonomy: dimensions of personality in the natural language and questionnaires," in L. A. Pervin, ed., *Handbook of personality: Theory and research*, New York, NY: Guilford Press, 1990, pp. 66–100.

Karlan, Dean and Jonathan Zinman, "Observing unobservables: Identifying information asymmetries with a consumer credit field experiment," *Econometrica*, November 2009, *77* (6), 1993—2008.

Lagarde, Mylene and Duane Blaauw, "Pro-social preferences and self-selection into jobs: Evidence from South African nurses," *Journal of Economic Behavior and Organization*, November 2014, *107* (A), 136–152.

Lazear, Edward P, "Performance Pay and Productivity," *American Economic Review*, December 2000, *90* (5), 1346–1361.

_ , "Teacher incentives," *Swedish Economic Policy Review*, 2003, *10* (3), 179–214.

Leaver, Clare, Renata Lemos, and Daniela Scur, "Measuring and explaining management in schools: New approaches using public data," Working Paper June 2019.

Lin, Winston, Donald P Green, and Alexander Coppock, "Standard operating procedures for Don Green's lab at Columbia," 2016.

Loyalka, Prashant, Sean Sylvia, Chengfang Liu, James Chu, and Yaojiang Shi, "Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement," *Journal of Labour Economics*, July 2019, *37* (3), 621–662.

Mbiti, Isaac, Mauricio Romero, and Youdi Schipper, "Designing Teacher Performance Pay Programs: Experimental Evidence from Tanzania," Working Paper 2018.

Muralidharan, Karthik and Venkatesh Sundararaman, "Teacher opinions on performance pay: Evidence from India," *Economics of Education Review*, 2011, *30*, 394–403.

_ **and** _ , "Teacher performance pay: Experimental evidence from India," *Journal of Political Economy*, February 2011, *119* (1), 39–77.

Niederle, Muriel and Lise Vesterlund, "Do women shy away from competition? Do men compete too much?," *Quarterly Journal of Economics*, 8 2007, *122* (3), 1067–1101.

Olken, Benjamin A, "Promises and perils of pre-analysis plans," *Journal of Economic Perspectives*, 2015, *29* (3), 61–80.

Rosenbaum, Paul R, *Design of Observational Studies*, New York: Springer-Verlag, 2010.

Rothstein, Jesse, "Teacher quality policy when supply matters," *American Economic Review*, 2015, *105* (1), 100–130.

Sheheryar, Banuri and Philip Keefer, "Pro-social motivation, effort and the call to public service," *European Economic Review*, April 2016, (83), 139–164.

Stallings, Jane A, Stephanie L Knight, and David Markham, "Using the Stallings Observation System to investigate time on task in four countries," World Bank Report No. 92558 2014.

Appendix A Supplemental figures

Figure A.1: Study profile

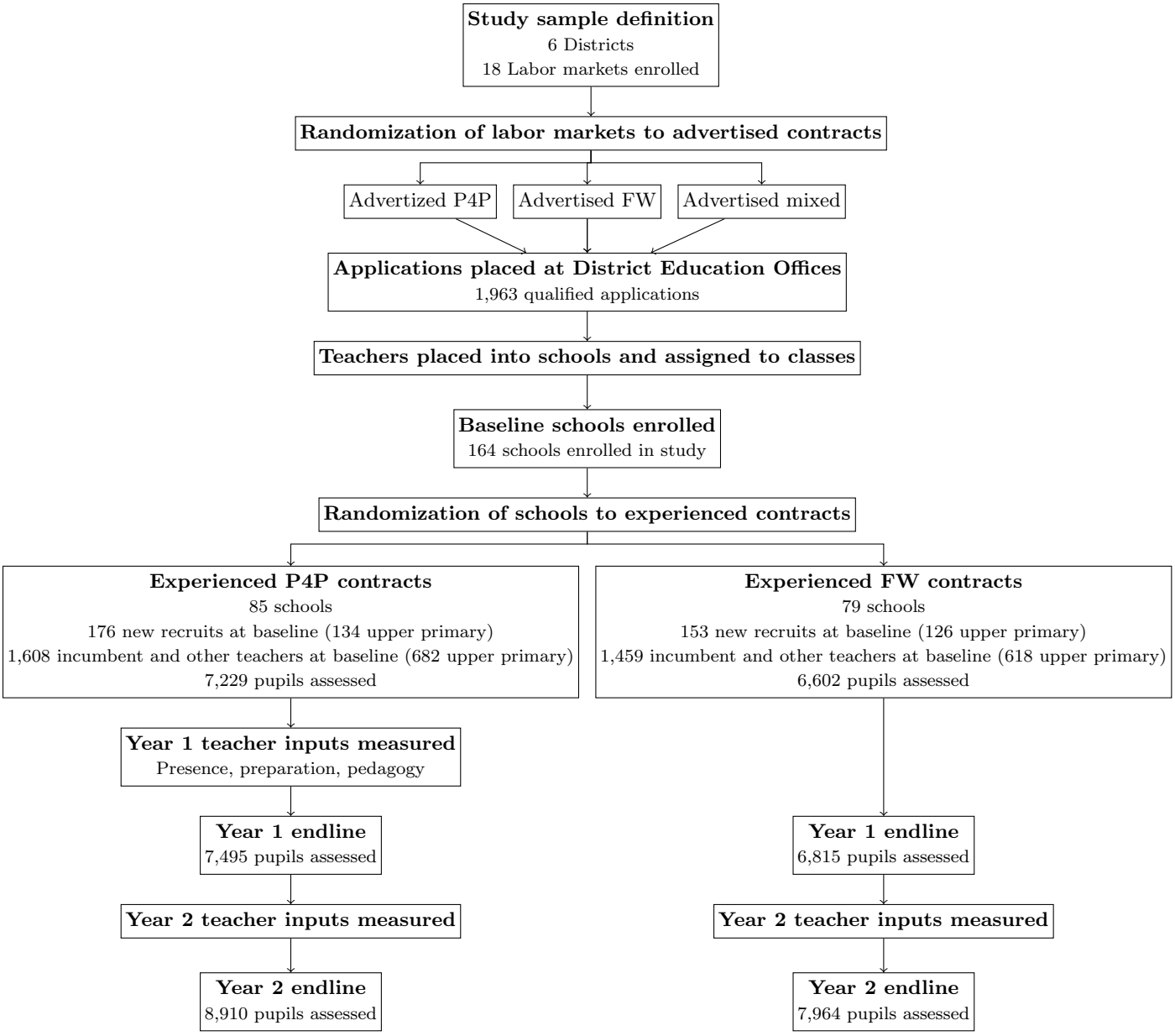


Figure A.2: Selection-out on baseline teacher skill, by experienced treatment

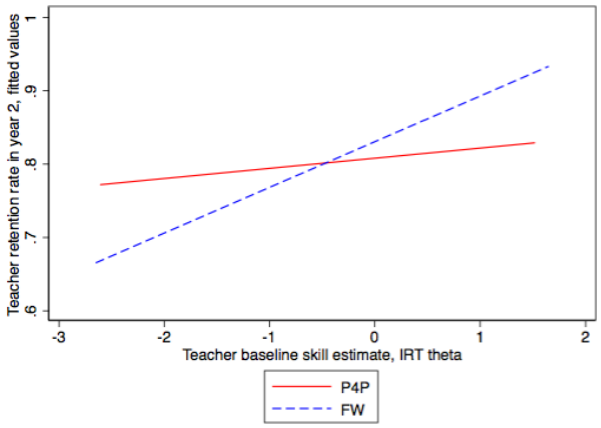
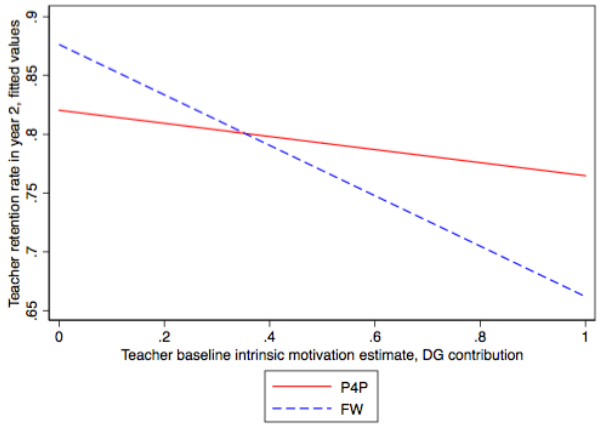


Figure A.3: Selection-out on baseline teacher intrinsic motivation, by experienced treatment



Appendix B Supplemental tables

Table B.1: Teacher endline survey responses

	Job satisfaction	Likelihood of leaving	Positive affect	Negative affect
Model A: Direct effects only				
Advertised P4P	-0.04	-0.07	-0.09	-0.00
	[-0.41, 0.53]	[-0.31, 0.11]	[-0.54, 0.38]	[-0.31, 0.46]
	(0.82)	(0.34)	(0.60)	(0.94)
Experienced P4P	0.05	-0.06	-0.01	0.09
	[-0.28, 0.39]	[-0.19, 0.08]	[-0.32, 0.31]	[-0.16, 0.37]
	(0.74)	(0.40)	(0.97)	(0.44)
Experienced P4P × Incumbent	0.00	0.04	0.04	-0.08
	[-0.52, 0.54]	[-0.15, 0.23]	[-0.53, 0.59]	[-0.57, 0.45]
	(0.98)	(0.62)	(0.82)	(0.69)
Model B: Interactions between advertised and experienced contracts				
Advertised P4P	-0.09	-0.01	0.02	-0.33
	[-0.63, 0.78]	[-0.33, 0.21]	[-0.65, 0.55]	[-0.84, 0.48]
	(0.71)	(0.91)	(0.89)	(0.20)
Experienced P4P	0.08	-0.07	-0.03	-0.24
	[-0.47, 0.60]	[-0.29, 0.17]	[-0.63, 0.54]	[-0.72, 0.22]
	(0.75)	(0.51)	(0.93)	(0.25)
Advertised P4P × Experienced P4P	0.12	-0.13	-0.24	0.68
	[-0.77, 0.90]	[-0.46, 0.18]	[-0.99, 0.42]	[0.01, 1.42]
	(0.75)	(0.32)	(0.44)	(0.02)
Experienced P4P × Incumbent	-0.02	0.05	0.06	0.26
	[-1.08, 1.10]	[-0.37, 0.42]	[-1.09, 1.10]	[-0.66, 1.25]
	(0.93)	(0.70)	(0.84)	(0.41)
Observations	1483	1492	1474	1447
FW recruit mean (SD)	5.42	0.26	0.31	0.00
	(0.90)	(0.44)	(0.93)	(0.99)
FW incumbent mean (SD)	5.26	0.29	-0.05	0.00
	(1.10)	(0.46)	(1.00)	(1.04)

Notes: The table presents estimated coefficients, 95 percent confidence intervals, and p-values. Means and standard deviations (in parentheses) of each outcome presented in footer, for recruits recruited and serving under fixed-wage contracts, and for incumbents serving under fixed-wage contracts.

Appendix C Theory

This appendix sets out a simple theoretical framework, adapted from Leaver et al. (2019), that closely mirrors the experimental design described in Section 2. We used this framework as a device to organize our thinking when choosing what hypotheses to test in our pre-analysis plan. Specifically, the framework helped us to make concrete terms frequently used in the literature (c.f. Dal Bó and Finan (2016)), such as 'teacher skill', 'teacher intrinsic motivation', 'selection', and 'incentives', and to develop distinct tests for the compositional (selection) and effort (incentive) margin effects on teacher performance. We did not view the framework as a means to deliver sharp predictions for one-tailed tests.

The model

We focus on an individual who has just completed teacher training, and who must decide whether to apply for a teaching post in a public school, or a job in a generic 'outside sector'.[31]

Preferences. The individual is risk neutral and cares about compensation w and effort e. Effort costs are sector-specific. The individual's payoff in the education sector is $w - (e^2 - \tau\, e)$, while her payoff in the outside sector is $w - e^2$. The parameter $\tau \geq 0$ captures the individual's *intrinsic motivation* to teach, and can be thought of as the realization of a random variable. The individual observes her realization τ perfectly, while (at the time of hiring) employers observe nothing.

Performance metrics. Irrespective of where the individual works, her effort generates a performance metric $m = e\,\theta + \varepsilon$. The parameter $\theta \geq 1$ is the individual's *ability*, and can also be thought of as the realization of a random variable. The individual observes her realization of θ perfectly, while (at the time of hiring) employers observe nothing. Draws of the error term ε are made from $U\,[\underline{\varepsilon}, \overline{\varepsilon}]$, and are independent across employments.

Compensation schemes. As described in the Study Design section, individuals belong to one of four subgroups, as shown in the 2x2 matrix below.

		Advertised	
		FW	P4P
Experienced	FW	a	b
	P4P	c	d

Different compensation schemes are available depending on advertised treatment status. In the *advertised P4P treatment*, individuals choose between: (i) an education contract of the form, $w^G + B$ if $m \geq \overline{m}$, or w^G otherwise; and (ii) an outside option of the form w^0 if $m \geq \underline{m}$, or 0 otherwise.

[31] Leaver et al. (2019) focus on a teacher who chooses between three alternatives: (i) accepting an offer of a job in a public school on a fixed wage contract, (ii) declining and applying for a job in a private school on a P4P contract, and (iii) declining and applying for a job in an outside sector on a (different) P4P contract.

In the *advertised FW treatment*, individuals choose between: (i) an education contract of the form w^F; and (ii) the same outside option. In our experiment, the bonus B was valued at RWF 100,000, and the fixed-wage contract exceeded the guaranteed income in the P4P contract by RWF 20,000 (i.e. $w^F - w^G = 20,000$).

Timing. The timing of the game is as follows.

1. Outside options and education contract offers are announced.
2. Nature chooses type (τ, θ).
3. Individuals observe their type (τ, θ), and choose which sector to appply to.
4. Employers hire (at random) from the set of applicants.
5. *Surprise* re-randomization occurs.
6. Individuals make effort choice e.
7. Individuals' performance metric m is realized, with $\varepsilon \sim U[\underline{\varepsilon}, \bar{\varepsilon}]$.
8. Compensation paid in line with (experienced) contract offers.

Numerical example To illustrate how predictions can be made using this framework, we draw on a numerical example. First, in terms of the compensation schemes, we assume that $w^O = 50$, $B = 40$, $w^G = 15$, $\underline{m} = 1$, and $\overline{m} = 4.5$ (as illustrated in Figure C.1). These five parameters, together with $\underline{\varepsilon} = -5$ and $\bar{\varepsilon} = 5$, pin down effort and occupational choices by a *given* (τ, θ)-type. If, in addition, we make assumptions concerning the distributions of τ and θ, then we can also make statements about the expected intrinsic motivation and expected ability of applicants, and the expected performance of placed recruits. Here, since our objective is primarily pedagogical, we go for the simplest case possible and assume that τ and θ are drawn independently from uniform distributions. Specifically, τ is drawn from $U[0, 10]$, and θ is drawn from $U[1, 5]$.

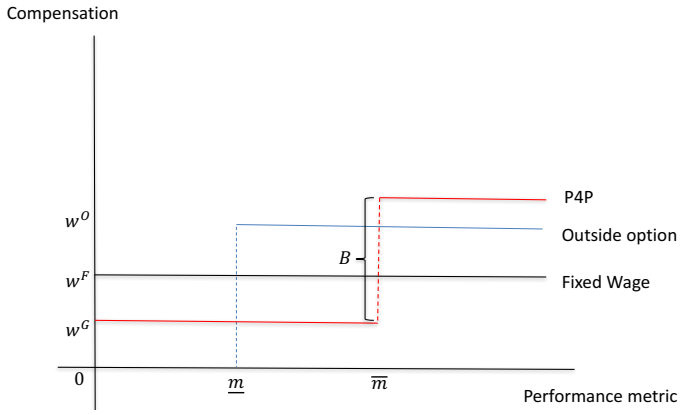
Analysis

As usual, we solve backwards, starting with effort choices.

Effort incentives Effort choices under the three compensation schemes are:

$$
\begin{aligned}
e^F &= \tau/2 \\
e^P &= \frac{\theta\, B}{2(\bar{\varepsilon} - \underline{\varepsilon})} + \tau/2 \\
e^O &= \frac{\theta\, w^O}{2(\bar{\varepsilon} - \underline{\varepsilon})},
\end{aligned}
$$

Figure C.1: Compensation schemes in the numerical example



where we have used the fact that ε is drawn from a uniform distribution. Intuitively, effort incentives are higher under P4P than under FW, i.e. $e^P > e^F$.

Supply-side selection. The individual applies for a teaching post advertised under P4P if, given her (τ, θ) type, she expects to receive a higher payoff teaching in a school on the P4P contract than working in the outside sector. We denote the set of such (τ, θ) types by $\mathcal{T}^P$. Similarly, the individual applies for a teaching post advertised under FW if, given her (τ, θ) type, she expects to receive a higher payoff teaching in a school on the FW contract than working in the outside sector. We denote the set of such (τ, θ) types by $\mathcal{T}^F$. Figure C.2 illustrates these sets for the numerical example. Note that the function $\tau^*(\theta)$ traces out motivational types who, given their ability, are just indifferent between applying to the education sector under advertised P4P and applying to the outside sector, i.e.:

$$\Pr\left[\theta e^P + \varepsilon > \overline{m}\right]\ B + w^G - (e^P)^2 + \tau^* e^P = \Pr\left[\theta e^O + \varepsilon > \underline{m}\right]\ w^O - (e^O)^2.$$

Similarly, the function $\tau^{**}(\theta)$ traces out motivational types who, given their ability, are just indifferent between applying to the education sector under advertised FW and applying to the outside sector, i.e.:

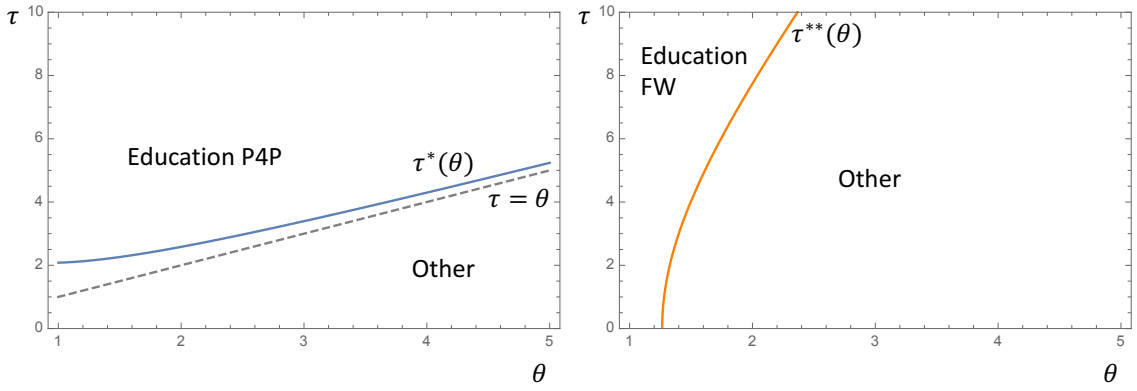
$$w^F - (e^F)^2 + \tau^{**} = \Pr\left[\theta e^O + \varepsilon > \underline{m}\right] \cdot w^O - (e^O)^2.$$

In the numerical example, we see a case of *positive selection on intrinsic motivation* and *negative selection on ability* under both the FW and P4P treatments. But there is *less* negative selection on ability under P4P than under FW.

Empirical implications

We used this theoretical framework when writing our pre-analysis plan to clarify *what* hypotheses to test. We summarize this process for Hypotheses I and VI below.

Figure C.2: Decision rules under alternative contract offer treatments



Hypothesis I: Advertised P4P induces differential application qualities. Define $1_{\{(\tau,\theta)\in\mathcal{T}^F\}}$ and $1_{\{(\tau,\theta)\in\mathcal{T}^P\}}$ as indicator functions for the application event in the advertised FW and P4P treatments respectively. The difference in expected intrinsic motivation and expected ability across the two advertised treatments, can be written as:

$$\mathrm{E}\left[\tau \cdot 1_{\{(\tau,\theta)\in\mathcal{T}^F\}}\right] - \mathrm{E}\left[\tau \cdot 1_{\{(\tau,\theta)\in\mathcal{T}^P\}}\right]$$

and

$$\mathrm{E}\left[\theta \cdot 1_{\{(\tau,\theta)\in\mathcal{T}^F\}}\right] - \mathrm{E}\left[\theta \cdot 1_{\{(\tau,\theta)\in\mathcal{T}^P\}}\right].$$

In the numerical example, both differences are negative: expected intrinsic motivation and expected ability are higher in the P4P treatment than in the FW treatment.

Hypothesis VI: Selection and incentive effects are apparent in the composite 4P performance metric. We start with the selection effect. Maintaining the assumption of no demand-side selection treatment effects, and using the decomposition in Leaver et al. (2019), we can write the difference in expected performance across sub-groups a and b (i.e. placed recruits who experienced FW) as:

$$\mathrm{E}[m^a] - \mathrm{E}[m^b] = \underbrace{\mathrm{E}\left[(\theta\, e^F - \theta\, e^F) \cdot 1_{\{(\tau,\theta)\in\mathcal{T}^F\}}\right]}_{\text{incentive effect} = 0} + \underbrace{\mathrm{E}\left[\theta\, e^F \cdot \left(1_{\{(\tau,\theta)\in\mathcal{T}^F\}} - 1_{\{(\tau,\theta)\in\mathcal{T}^P\}}\right)\right]}_{\text{selection effect}}.$$

Similarly, the difference in expected performance across sub-groups c and d (i.e. placed recruits who experienced P4P) can be written as:

$$\mathrm{E}[m^c] - \mathrm{E}[m^d] = \underbrace{\mathrm{E}\left[(\theta\, e^P - \theta\, e^P) \cdot 1_{\{(\tau,\theta)\in\mathcal{T}^F\}}\right]}_{\text{incentive effect} = 0} + \underbrace{\mathrm{E}\left[\theta\, e^P \cdot \left(1_{\{(\tau,\theta)\in\mathcal{T}^F\}} - 1_{\{(\tau,\theta)\in\mathcal{T}^P\}}\right)\right]}_{\text{selection effect}}.$$

In the numerical example, both differences are negative, and the second is larger than the first.

Turning to the incentive effect, we can write the difference in expected performance across sub-groups a and c (i.e. placed recruits who applied under advertised FW) as:

$$\mathrm{E}[m^a] - \mathrm{E}[m^c] = \underbrace{\mathrm{E}\left[(\theta\, e^F - \theta\, e^P) \cdot 1_{\{(\tau,\theta)\in\mathcal{T}^F\}}\right]}_{\text{incentive effect}} + \underbrace{\mathrm{E}\left[\theta\, e^F \cdot \left(1_{\{(\tau,\theta)\in\mathcal{T}^F\}} - 1_{\{(\tau,\theta)\in\mathcal{T}^F\}}\right)\right]}_{\text{selection effect=0}}.$$

Similarly, the difference in expected performance across sub-groups b and d (i.e. placed recruits who applied under advertised P4P) can be written as:

$$\mathrm{E}[m^b] - \mathrm{E}[m^d] = \underbrace{\mathrm{E}\left[(\theta\, e^F - \theta\, e^P) \cdot 1_{\{(\tau,\theta)\in\mathcal{T}^P\}}\right]}_{\text{incentive effect}} + \underbrace{\mathrm{E}\left[\theta\, e^P \cdot \left(1_{\{(\tau,\theta)\in\mathcal{T}^P\}} - 1_{\{(\tau,\theta)\in\mathcal{T}^P\}}\right)\right]}_{\text{selection effect=0}}.$$

In the numerical example, both differences are negative, and the second is larger than the first. Hypothesis IV and V focus on one component of the performance metric—student performance—and follow from the above.