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[§¶]

December 31, 2019

Abstract

This paper reports results of a two-tiered experiment designed to separately identify the selection and effort margins of pay-for-performance (P4P). At the recruitment stage, teacher labor markets were randomly assigned to a pay-for-percentile or fixed-wage contract. Once recruits had been placed, an unexpected, incentive-compatible, school-level re-randomization was performed, so that some teachers who applied for a fixed-wage contract ended up being paid by P4P, and vice versa. Pooling across two years of exposure, study results show positive effects of experienced P4P contracts and rule out meaningful negative recruitment effects of P4P on pupil learning. By year two of the study, P4P increased pupil learning per grade by 0.21 standard deviations. One quarter of this impact can be attributed to favorable selection on unobserved traits at the recruitment stage, with the remainder arising from increased effort, including improved teacher presence and pedagogy.

JEL codes: C93, I21, J45, M52, O15.
Keywords: pay-for-performance, selection, incentives, teachers, field experiment.

[*]Blavatnik School of Government, University of Oxford, Toulouse School of Economics, CEPR, and RISE (clare.leaver@bsg.ox.ac.uk)

[†]World Bank Development Research Group, BREAD, and IZA (oozier@worldbank.org)

[‡]School of International Development, University of East Anglia, IZA, RISE, and EGAP (p.serneels@uea.ac.uk)

[§]McCourt School of Public Policy, Georgetown University, CGD, and IGC (az332@georgetown.edu)

[¶]We thank counterparts at REB and MINEDUC for advice and collaboration, and David Johnson for help with the design of student and teacher assessments. We are grateful to Katherine Casey, Jasper Cooper, Ernesto Dal Bó, Erika Deserranno, David Evans, Dean Eckles, Frederico Finan, James Habyarimana, Caroline Hoxby, Macartan Humphreys, Pamela Jakiela, Julien Labonne, David McKenzie, Ben Olken, Berk Özler, Cyrus Samii, Kunal Sen, Martin Williams, and audiences at BREAD, DfID, EDI, NBER, SIOE, and SREE for helpful comments. IPA staff members Kris Cox, Stephanie De Mel, Olive Karekezi Kemirembe, Doug Kirke-Smith, Emmanuel Musafiri, and Phillip Okull, and research assistants Claire Cullen, Robbie Dean, Ali Hamza, Gerald Ipapa, and Saahil Karpe all provided excellent support. Financial support was provided by the U.K. Department for International Development (DfID) via the International Growth Centre and the Economic Development and Institutions Programme, by Oxford University's John Fell Fund, and by the World Bank's SIEF and REACH trust funds. We received IRB approval from the Rwanda National Ethics Committee and from Innovations for Poverty Action. This study is registered as AEA RCT Registry ID AEARCTR-0002565. The findings in this paper 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.

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 a sector such as education where people—that is, human rather than physical resources—play a key role. Effective teachers generate private returns for students through learning gains, educational attainment, and higher earnings (Chetty et al., 2014a,b), as well as social returns through improved labor-market skills that drive economic growth (Hanushek and Woessmann, 2012). And yet in varying contexts around the world, governments struggle to maintain a skilled and motivated teacher workforce (Bold et al., 2017).

One policy option in this context is *pay-for-performance*. These compensation schemes typically reward teacher inputs such as presence and conduct in the classroom, teacher value-added based on student learning, or both (see, e.g., Muralidharan and Sundararaman, 2011b). In principle, they can address the difficulty of screening for teacher quality ex ante (Staiger and Rockoff, 2010), as well as the limited oversight of teachers on the job (Chaudhury et al., 2006).

Yet pay-for-performance divides opinion. Critics, drawing upon public administration, social psychology, and behavioral economics, argue that pay-for-performance could dampen the effort of workers (Bénabou and Tirole, 2003; Deci and Ryan, 1985; Krepps, 1997). Concerns are that such schemes may: recruit the wrong types, individuals who are "in it for the money"; lower effort by eroding intrinsic motivation; and fail to retain the right types because good teachers become de-motivated and quit. By contrast, proponents point to classic contract theory (Lazear, 2003; Rothstein, 2015) and evidence from private-sector jobs with readily measurable output (Lazear, 2000) to argue that pay-for-performance will have positive effects on both compositional and effort margins. Under this view, such schemes: recruit the right types, individuals who anticipate performing well in the classroom; raise effort by strengthening extrinsic motivation; and retain the right types because good teachers feel rewarded and stay put.

This paper conducts the first prospective, randomized controlled trial designed to identify both the compositional and effort margins of pay-for-performance. A novel, two-tiered experiment separately identifies these effects. This is combined with detailed data on applicants to jobs, the skills and motivations of actual hires, and their performance over two years on the job, to evaluate the effects of pay-for-performance on the recruitment, effort, and retention of civil servant teachers.

At the center of this study is a pay-for-performance (hereafter P4P) contract, designed jointly with the Rwanda Education Board and Ministry of Education. Building on extensive consultations and a pilot year, this P4P contract rewards the top 20 percent of teachers with extra pay 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). The measure of learning used was based on a *pay-for-percentile* scheme that makes student performance at all levels relevant to teacher rewards (Barlevy and Neal, 2012). The tournament nature of this contract allows us to compare it to a fixed-wage (hereafter FW) contract that is equal in expected payout.

Our two-tiered experiment first randomly assigns labor markets to either P4P or 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. The first stage was undertaken during recruitment for teacher placements for the 2016 school year. Teacher labor markets are defined at the district by subject-family level. We conducted the experiment in six districts (18 labor markets) which, together, cover more than half the upper-primary teacher hiring lines for the 2016 school year. We recruited into the study all primary schools that received such a teacher to fill an upper-primary teaching post (a total of 164 schools). The second stage was undertaken once 2016 teacher placements had been finalized. Here, we randomly re-assigned each of these 164 study schools in their entirety to either P4P or FW contracts; all teachers who taught core-curricular classes to upper-primary students, including both newly placed recruits and incumbents, 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. As advertised at the time of recruitment, 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 three main findings are as follows. First, on recruitment, advertised P4P contracts did not change the distribution of measured teacher skill either among applicants in general or among new hires in particular. This is estimated sufficiently precisely to rule out even small negative effects of P4P on measured skills. Advertised P4P contracts did, however, select teachers who contributed less in a framed Dictator Game played at baseline to measure intrinsic motivation. In spite of this,

teachers recruited under P4P were at least as effective in promoting learning as were those recruited under FW (holding experienced contracts constant), delivering in point estimate terms an additional 0.05 standard deviations of learning per year by their second year on the job.[1]

Second, in terms of incentivizing effort, placed teachers working under P4P contracts elicited better performance from their students than teachers working under FW contracts (holding advertised contracts constant). They delivered an additional 0.11 standard deviations of learning per year on average over the two years of the study, rising to 0.16 standard deviations in their second year on the job. There is no evidence of a differential impact of experienced contracts by type of advertisement.

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 8 percentage points higher among recruits who experienced the P4P contract compared to recruits who experienced the FW contract. This is 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 0.10 points, as measured on a 4-point scale.

Third, on retention, teachers working under P4P contracts were no more likely to quit during the two years of the study than teachers working under FW contracts. There was also no evidence of differential selection-out on baseline teacher characteristics by experienced contract, either in terms of skills or measured motivation. 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. By the second year of the study, we estimate the total effect of P4P to be 0.21 standard deviations of pupil learning per year. 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

[1]This modest positive effect on recruitment is supported by supplementary analysis: a substantially larger OLS estimate of the advertised P4P effect, and first-order stochastic dominance of the distribution of teacher value added under advertised P4P compared to advertised FW.

incentivized dimensions such as teacher presence and pedagogy.

Our findings bring new experimental results on pay-for-performance to the literature on the recruitment of civil servants in low- and middle-income countries. Existing papers have examined the impact of advertising higher *unconditional* salaries and career-track motivations, with mixed results. In Mexico, Dal Bó et al. (2013) find that higher base salaries attracted both skilled and motivated applicants for civil service jobs. In Uganda, Deserranno (2019) finds that the expectation of higher earnings discouraged pro-social applicants for village promoter roles, resulting in lower effort and retention. And in Zambia, Ashraf et al. (forthcoming) find that emphasis on career-track motivations for community health work, while attracting some applicants who were less pro-social, resulted in hires of equal pro-sociality and greater talent overall, leading to improvements in a range of health outcomes. By studying pay-for-performance and by separately manipulating advertised and experienced contracts, we add evidence on the compositional and effort margins of a different, and widely debated, compensation policy for civil servants.

How the teaching workforce changes in response to pay-for-performance is of interest in high-income contexts as well. In the US, there is a large (but chiefly observational) literature on the impact of compensation on who enters and leaves the teaching workforce. Well-known studies have simulated the consequences of dismissal policies (Chetty et al., 2014b; Neal, 2011; Rothstein, 2015) or examined the role of teachers' outside options in labor supply (Chingos and West, 2012). Recent work has examined the District of Columbia's teacher evaluation system, where financial incentives are linked to 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) confirm that these 'quitters' were replaced by higher-performers. In Wisconsin, a reform permitted approximately half of the state's school districts to introduce flexible salary schemes that allow pay to vary with performance. In that setting, Biasi (2019) finds that high-value-added teachers were more likely to move to districts with flexible pay, and were less likely to quit, than their low-value-added counterparts. Our prospective, experimental study of pay-for-performance contributes to this literature methodologically but also substantively since the Rwandan labor market shares important features with high-income contexts.[2]

[2]Notably, there is no public sector pay premium in Rwanda, which is unusual for a low-income country and more typical of high-income countries (Finan et al., 2017). The 2017 Rwanda Labour

While our paper is not the first on the broader topic of incentive-based contracts for teachers,[3] we go to some length to incorporate two features thought to be essential to policy implementation at scale. One is that the structure of the incentive does not unfairly disadvantage any particular group (Barlevy and Neal, 2012); the other is that the incentive should not be inappropriately narrow (Stecher et al., 2018). We address the first issue by using a measure of learning based on a pay-for-percentile scheme that makes student performance at all levels relevant to teacher rewards, and the second by combining this with measures of teachers' inputs into the classroom to create a broad, composite metric. There is a small but growing literature studying pay-for-percentile schemes in education: Loyalka et al. (2019) in China, Gilligan et al. (forthcoming) in Uganda, and Mbiti et al. (2019) in Tanzania. Our contribution is to compare the effectiveness of contracts, P4P versus FW, that are based on a composite metric and are budget neutral in salary.

A final, 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, policy 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).[4] 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 on the basis of their realized statistical significance. The spirit of this approach is similar to recent work by Anderson and Magruder (2017) and Fafchamps and Labonne (2017).[5] For an experimental study in which one important dimension of variation occurs at the labor-market level,

Force Survey includes a small sample of recent Teacher Training College graduates (aged below age 30). Of these, 37 percent were in teaching jobs earning an average monthly salary of 43,431 RWF , while 15 percent were in non-teaching jobs earning a higher average monthly salary of 56,347 RWF —a *private sector* premium of close to 30 percent.

[3]See, e.g., Imberman and Lovenheim (2015) and Jackson et al. (2014) who provide a review.

[4]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.

[5]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.

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.

In the remainder of the paper, Section 2 and 3 describe the study design and data, Section 4 and 5 report and discuss the results, and Section 6 concludes.

2 Study design

2.1 Setting

The first tier of the study took place during the actual recruitment for civil service teaching jobs in upper primary in six districts of Rwanda in 2016.[6] To apply for a civil service teaching job an individual needs to hold a Teacher Training College (TTC) degree. Eligibility is further defined by specialization. The Rwanda Education Board (REB) confirmed that districts solicit applications at the district-by-subject-family level, aggregating curricula subjects into three 'families' that correspond to the degree types issued by TTCs: math and science (TMS); modern languages (TML); and social studies (TSS). Districts invite applications between November and December, for the academic year beginning in late January/early February. Individuals keen to teach in a particular district submit one application and are then considered for all eligible teaching posts in that district in that hiring round. It is difficult to apply to many districts, and the majority of applicants apply to only one district.[7]

Given this institutional setting, one can think of district-by-subject-family pairs as distinct *labor markets*. There are 18 such markets in our study.[8] This is a small number in terms of statistical power (as we address below) but not from a system-scale perspective. The study covers more than 600 hiring lines constituting over 60 percent of the country's planned recruitment in 2016. Importantly, it is not a foregone conclusion that TTC graduates will apply for these civil service teaching jobs. Data from the 2017 Rwanda Labour Force Survey indicate that only 37 percent of recent TTC graduates were in teaching jobs, with 15 percent in non-teaching,

[6] Upper primary refers to grades 4, 5, and 6; schools typically include grades 1 through 6.

[7] Applying to multiple districts is logistically costly, since each district requires its own exam.

[8] Inference based on asymptotics could easily be invalid with 18 randomizable markets. We address this risk by committing to randomization inference for all aspects of statistical testing.

salaried employment. The recent graduates in the outside sector earned a premium of close to 30 percent, making occupational choice after TTC a meaningful decision.

2.2 Experiment

Contract structure The experiment was built around the comparison of two contracts paying a bonus on top of teacher salaries in each of the 2016 and 2017 school years, and was managed by Innovations for Poverty Action (IPA) in coordination with REB. The first of these was a pay-for-performance (P4P) contract, which paid 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.[9] This metric equally weighted student learning alongside three measures of teachers' inputs into the classroom (presence, lesson preparation, and observed pedagogy). The measure of learning was based on a pay-for-percentile scheme that makes student performance at all levels relevant to teacher rewards (Barlevy and Neal, 2012).[10] The 2016 performance award was conditional on remaining in post during the entire 2016 school year, and was to be paid early in 2017. Likewise, the 2017 performance award was conditional on remaining in post during the entire 2017 school year, and was to be paid early in 2018. The second was a fixed-wage (FW) contract that paid RWF 20,000 to all upper-primary teachers. This bonus was paid at the same time as the performance award in the P4P contract.

Design overview The design, summarized visually in Figure A.1, draws on a two-tiered experiment, as used elsewhere (see Karlan and Zinman (2009), Ashraf et al. (2010), and Cohen and Dupas (2010) in credit-market and public-health contexts). Both tiers employ the contract variation described above.

Potential applicants, not all of whom were observed, were assigned to either advertised FW or advertised P4P contracts, depending on the labor market in which they resided. Those who actually applied, and were placed into schools, fall into one of the four groups summarized in Figure 1. For example, group a denotes teachers

[9]The exchange rate on January 1, 2016 was 734 RWF to 1 USD, so the RWF 100,000 bonus was worth roughly 136 USD.

[10]Student learning contributed to an individual teacher's score via percentiles within student-based brackets so that a teacher with a particular mix of low-performing and high-performing students was, in effect, competing with other teachers with similar mixes of students. Full details are provided in the pre-analysis plan. The data used to construct this measure, and the measures of teachers' inputs, are described in Section 3.3 and 3.4 respectively.

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 who were then placed in schools re-randomized to P4P contracts. Under this experimental design, 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, whereas comparisons along the diagonal of a–d are informative about the total effect of such contracts, along both margins.

Figure 1: Treatment groups among recruits placed in study schools

		Advertised	
		FW	P4P
Experienced	FW	a	b
	P4P	c	d

First tier randomization: Advertised contracts Our aim in the first tier was to randomize the 18 distinct labor markets to contracts, 'treating' all potential applicants in a given market so that we could detect the supply-side response to a particular contract. The result of the randomized assignment is that 7 of these labor markets can be thought of as being in a 'P4P only' advertised treatment, 7 in a 'FW only' advertised treatment, and 4 in a 'Mixed' advertised treatment.[11] Empirically, we consider the Mixed treatment as a separate arm; we estimate a corresponding advertisement effect 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. 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 TTCs to promote the interventions.

[11]This randomization was performed in MATLAB by the authors. The Mixed advertised treatment arose due to logistical challenges detailed in the pre-analysis plan: the first-tier randomization was carried out at the level of the subject rather than the subject-family. An example of a district-by-subject-family assigned to the Mixed treatment is Ngoma-TML. An individual living in Ngoma with a TML qualification could have applied for an advertised Ngoma post in English on a FW contract, or an advertised Ngoma post in Kinyarwanda on a P4P contract. In contrast, Kirehe-TML is in the P4P only treatment. So someone in Kirehe with a TML qualification could have applied for either an English or Kinyarwanda post, but both would have been on a P4P contract.

These job fairs were advertised through WhatsApp networks of TTC graduates. All 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. Applications were then submitted in December 2015. In January 2016, all districts held screening examinations for potential candidates. Successful candidates were placed into schools by districts during February–March 2016, and were then assigned to particular grades, subjects, and streams by their head teachers.

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. Following a full baseline survey in February 2016, schools were randomly assigned to either P4P or FW. Of the 164 schools in the second tier of the experiment, 85 were assigned to P4P and 79 were assigned to FW contracts.

All upper-primary teachers—placed recruit or incumbent—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 behavior. 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, surprise effects, disappointment or otherwise, may still be present in on-the-job performance. When testing hypotheses relating to student learning, we include a secondary specification with an interaction term to allow the estimated impact of experienced P4P to differ by advertised treatment. We also explore whether

surprise effects are evident in either retention or job satisfaction. We find no evidence for any surprise effect.

2.3 Hypotheses

Pre-commitment to an analytical approach can forestall p-hacking, but requires clear specification of both what to test and how to test it; this presents an opportunity, as we now discuss. A theoretical model, discussed briefly below, and included in our pre-analysis plan, guides our choice of *what* hypotheses to test. However, exactly *how* to test these hypotheses in a way that maximizes statistical power is not fully determined by theory, as statistical power may depend on features of the data that could not be known in advance: the distribution of outcomes, their relationships with possible baseline predictors, and so on. We used blinded data to help decide how to test the hypotheses. In what follows we first briefly describe the theoretical model, and then discuss our statistical approach.

Theory The model considers a fresh graduate from teacher training who decides whether to apply for a state school teaching post, or a job in another sector (a composite 'outside sector'). The risk neutral individual cares about compensation w and effort e. Her payoff is sector specific: in teaching it is $w - (e^2 - \tau\, e)$, while in the outside sector it is $w - e^2$. The parameter $\tau \geq 0$ captures the individual's *intrinsic motivation* to teach, which is perfectly observed by the individual herself but not by the employer at the time of hiring.[12] Effort generates $m = e\,\theta + \varepsilon$, where $\theta \geq 1$ represents her *ability*, which is also private information at the time of hiring. Compensation corresponds to one of the four cells in Figure 1. The timing is as follows. Teacher vacancies are advertised as either P4P or FW. The individual, of type (τ, θ), applies either to a teaching job or to an outside job. Employers hire, at random, from the set of (τ, θ) types that apply. Thereafter, contracts are re-randomised. If the individual applies to, and is placed in a school, she learns about her experienced contract and chooses her effort level, which results in performance m at the end of the year. Compensation is paid according to the experienced contract.

This model leads to the following hypotheses, as set out in our pre-analysis plan:

[12] See Delfgaauw and Dur (2007) for a related approach to modeling differential worker motivation across sectors.

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 supply-side selection-in of higher- (or lower-)performing teachers, as measured by the learning outcomes of their students;

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

VI. These selection and incentive effects are apparent in the composite performance metric.

The model predicts that the set of (τ, θ) types preferring a teaching job advertised under P4P to a job in the outside sector is different from the set of types preferring a teaching job advertised under FW to a job in the outside sector. This gives Hypothesis I. Since the model abstracts from labor demand effects (by assuming employers hire at random from the set of (τ, θ) types that apply), this prediction simply maps through to placed recruits; i.e. to Hypothesis II via θ, Hypothesis III via τ, and Hypothesis IV to VI via the effect of θ and τ on performance.[13] The model also predicts that any given (τ, θ) type who applies to, and is placed in, a teaching job will exert more effort under experienced P4P than experienced FW. This gives Hypothesis V and VI via the effect of e on performance.

Analysis of blinded data Combining several previously-known insights, we used blinded data to maximize statistical power for our main hypothesis tests.

The first insights, pertaining to simulation, are due to Humphreys et al. (2013) and Olken (2015). Researchers can use actual outcome data with the treatment variable scrambled or removed to estimate specifications in 'mock' data. This permits navigation of an otherwise intractable 'analysis tree'. They can also improve

[13] When mapping the theory to our empirical context, we distinguish between these hypotheses for several reasons. First, as we discuss in Appendix C, the hiring rule used could mean that we observe a different advertised treatment effect among the sub-sample of placed recruits versus the sample of applicants. Second, we have better data for placed recruits because we were able to administer detailed survey instruments to this well-defined sub-sample. Third, for the sub-sample of placed recruits we can identify the advertised treatment effect from student learning outcomes, avoiding the use of proxies for (τ, θ).

statistical power by simulating treatment effects and choosing the specification that minimizes the standard error. Without true treatment assignments, the influence of any decision over eventual treatment effect estimates is unknown; thus these benefits are garnered without risk of p-hacking.[14]

The second set of insights pertain to randomization inference. Since the market-level randomization in our study involves 18 randomizable units, asymptotic inference is unsuitable, so we use randomization inference. It is known that any scalar function of treatment and comparison groups is a statistic upon which a (correctly-sized) randomization-inference-based test of the sharp null hypothesis could be built, but also that such statistics may vary in their statistical power in the face of any particular alternative hypothesis (Imbens and Rubin, 2015). We anticipated that, even with correctly-sized tests, the market-level portion of our design may present relatively low statistical power. Consequently, we conducted blinded analysis to choose, on the basis of statistical power, among testing approaches for several hypotheses: Hypothesis I, and a common framework for Hypotheses IV and V.[15]

Hypothesis I is the test of whether applicants to different contracts vary in their TTC scores. Blinded analysis, in which we simulated additive treatment effects and calculated the statistical power under different approaches, suggested that ordinary least squares regression (OLS) would provide markedly lower statistical power than would a Kolmogorov-Smirnov (KS) test of the equality of two distributions. Across a range of simulations, we found the KS test to have between one and four times the power of OLS. We therefore committed to the KS test (over OLS and two other alternatives) as our primary test of this hypothesis. This prediction is borne out in Appendix Table C.1.[16]

Hypotheses IV and V relate to the effects of advertised and experienced contracts on student test scores. Here, with the re-randomization taking place at the

[14]This would not be true if, for example, an outcome in question was known to have different support as a function of treatment, allowing the 'blinded' researcher to infer treatment from the outcome variable. For our blinded pre-analysis, we only consider outcomes (TTC score, and student test scores) that are nearly continuously distributed and which we believe are likely to have the same support in all study arms. To make this analysis possible, we drew inspiration from Fafchamps and Labonne (2017), who suggest dividing labor within a research team. In our case, IPA oversaw the data-blinding process. Results of the blinded analysis (for which IPA certified that we used only blinded data) are in our pre-analysis plan. Our RCT registry entry (AEARCTR-0002565) is accompanied by IPA's letter specifying the date after which treatment was unblinded.

[15]Hypotheses II and III employ data that our team collected, so did not have power concerns associated with them; Hypothesis VI offered fewer degrees of freedom.

[16]The confidence interval for the KS test is roughly half the width of the corresponding OLS confidence interval: a gain in precision commensurate with more than tripling the sample size.

school level, we had many more possible specifications to choose from. We examined 14 specifications (modeling random effects or fixed effects at different levels), and committed to one with the highest power. Simulations suggested that this would produce a 20 to 25 percent narrower confidence interval than in a simple benchmark specification. The accuracy of this prediction is borne out by comparing Table 3 to Appendix Table A.3.[17]

On the basis of this theory and analysis of blinded data, we settled on six primary tests: an outcome, a sample, a specification and associated test statistic, and an inference procedure for each of Hypotheses I-VI, as set out in Appendix Table A.1. We also included a small number of secondary tests based on different outcomes, samples, and/or specifications. In Section 4, we report results for every primary test; secondary tests are in Section 4 or in an appendix. To aid interpretation, we also include a small amount of supplementary analysis that was not discussed in the pre-analysis plan—e.g. estimates from a teacher value added model—but are cautious and make clear when this is post-hoc.

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. 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.[18] Of the 2,185 applications, 1,963 come from candidates with a TTC degree—we term these *qualified* since, at least in principle,

[17] Our pre-committed random effects model yields a confidence interval of roughly 81 percent the width of the analogous OLS confidence interval, commensurate with the power gain from increasing the sample size by half.

[18] These data were obtained from the six district offices and represent a census of applications for the new posts across these districts.

a TTC degree is required for the placements at stake. In the table, we present TTC scores, genders, and ages—the other observed CV characteristics—for all qualified applicants. Besides these two demographic variables, TTC scores are the only consistently measured characteristics of all applicants. 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; 62 percent of qualified applicants complete only one application.

Table 1: Application characteristics, by district

	Gatsibo	Kayonza	Kirehe	Ngoma	Nyagatare	Rwamagana	All
Applications	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
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

3.2 Teacher attributes

We visited the enrolled schools at baseline in February 2016, and collected data using surveys, and 'lab-in-the-field' instruments. School surveys were administered to head teachers or their deputies, 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 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, earnings, and other characteristics.

The 'lab-in-the-field' instruments were administered to the same set of teachers, and were intended to measure the two characteristics introduced in the theory: intrinsic motivation and ability. In the model, more intrinsically motivated teachers derive a higher benefit (or lower cost) from their efforts to promote learning. To capture this idea of other-regarding preferences towards students, taking inspiration from the work of Ashraf et al. (2014), we used a framed version of the *Dictator*

Game.[19] Teachers were given 2,000 Rwandan Francs (RWF) and asked how much of this money they wished to allocate to the provision of school supply packets for 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. Teachers could decide to allocate any amount, from zero to all 2,000 RWF, which would supply ten randomly chosen students with a packet.

We also asked teachers to undertake a *Grading Task* which measured their mastery of the curriculum in the main subject that they teach.[20] 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. Performance on this task was used to estimate a measure of teacher skill based on a two-parameter item response theory (IRT) model.

3.3 Student learning

Student learning was measured in three rounds of assessment: baseline, the end of the 2016 school year, and the end of the 2017 school year (indexed by $r = 0, 1, 2$). 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 experienced P4P arm for purposes of performance awards.

Working with the Ministry of Education, we developed comprehensive subject- and grade-specific, competency-based assessments for grades 4, 5 and 6. These assessments were based on the new Rwanda national curriculum and covered the five core subjects: Kinyarwanda, English, Mathematics, Sciences, and Social Studies. There was one assessment per grade-subject, with students at the beginning of the year being assessed on the prior year's material.[21] 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.[22] In each round, we randomly sampled a subset of students from each grade to take the test. In Year 1, both

[19] Previous work demonstrates the reliability of the Dictator Game (Eckel and Grossman, 1996) as a measure of other-regarding preferences that relate to intrinsic motivation, see for instance: Banuri and Keefer (2016); Brock et al. (2016); Deserranno (2019).

[20] See Bold et al. (2017) who use a similar approach to assess teacher content knowledge.

[21] A special grade 3 assessment was developed to assess grade 4 students at baseline.

[22] In piloting, all student tests were administered in English but we found that grade 4 students had not yet received sufficient English instruction. Grade 4 tests were therefore translated and administered in Kinyarwanda throughout the study.

baseline and endline student samples were drawn from the official school register of enrolled students compiled by the head teacher 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.

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.[23]

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.[24] 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 IRT model. 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 several dimensions of 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. This composite metric is based on three teacher input measures

[23]Consequently, the number of pupils assessed in Year 2 who have also been assessed in Year 1 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.

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

(presence, lesson preparation, and observed pedagogy), and one output measure (pupil learning)—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.[25] FW schools also received an unannounced visit in Year 2, at the same time as the P4P schools. Table A.2 shows summary statistics for each of these three input measures over the three rounds of the study.

Presence is defined as the fraction of spot-check days that the teacher is present at the start of the school day. 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 teachers' 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 guidelines.[26] Specifically, SEOs visited schools and provided teachers with a template to help prepare three key components of a lesson—the lesson objective, the instructional activities, and the types of assessment to be used. A 'hands-on' session then enabled teachers to practice writing lesson plans using this template before incorporating it into 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 did not have a lesson plan on a randomly chosen audit day, we used the presence of such a lesson plan as a summary measure in both the incentivized contracts and as an outcome for analysis.

Pedagogy is defined as the practices and methods that teachers use in order

[25]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 were shown videos recorded during pilot visits. 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.

[26]To isolate the effects of pay-for-performance pay, training was kept to a minimum and focused on how teachers could meet the targeted metrics.

to impact student learning. We collaborated with both the Ministry of Education and REB 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. Our rubric was 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.[27] Based on these observations and a detailed rubric, the observer provided a subjective score, on a scale from zero to three, of four components of the lesson: communication of lesson objectives, delivery of material, use of assessment, and student engagement.[28] The teacher's incentivized score, as well the measure of pedagogy used in our analysis, is defined as the average of these ratings across the four domains.

3.5 Balance

We use the baseline data described in this section to check whether the second-tier randomization produced an appropriately 'balanced' experienced treatment assignment. Table 2 confirms that across a wide range of school, teacher, and student characteristics there are no statistically significant differences in means between the experienced P4P and FW treatment arms.[29]

4 Results

Our two-tiered experiment allows us to estimate impacts of pay-for-performance on the type of individuals applying to, and being placed in, primary teaching posts (the compositional margin), and on the activities undertaken by these new recruits (the effort margin). We report these results in Sections 4.1 and 4.2 respectively. Of course, the long-run effects of pay-for-performance will depend not only on selection-in, but also selection-*out*, as well as the dynamics of the behavioral response on the part of teachers who stay. We address these dynamic issues in Section 4.3, and postpone a substantive discussion of results until Section 5. All statistical tests are

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

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

[29] Since the teacher inputs described in Section 3.4 were collected *after* the second-tier randomization, they are not included in Table 2. See instead Appendix Table A.2.

Table 2: Baseline characteristics and balance of experienced P4P assignment

	Control mean [St. Dev.]	Experienced P4P (p-value)	Obs.
Panel A. School attributes			
Number of streams	9.99	-0.10	164
	[4.48]	(0.881)	
Number of teachers	20.39	0.55	164
	[8.51]	(0.729)	
Number of new recruits	1.86	0.13	164
	[1.25]	(0.496)	
Number of students	410.06	1.42	164
	[206.71]	(0.985)	
Share female students	0.58	0.00	164
	[0.09]	(0.777)	
Panel B. Upper-primary teacher attributes			
Female	0.38	-0.03	225
	[0.49]	(0.646)	
Age	25.98	-0.18	249
	[4.17]	(0.742)	
DG share sent	0.28	-0.03	238
	[0.33]	(0.524)	
Grading task score	-0.24	0.12	238
	[0.93]	(0.276)	
Panel C. Pupil learning assessments			
English	-0.00	0.04	13826
	[1.00]	(0.551)	
Kinyarwanda	-0.00	0.05	13831
	[1.00]	(0.292)	
Mathematics	0.00	-0.00	13826
	[1.00]	(0.950)	
Science	-0.00	0.03	13829
	[1.00]	(0.607)	
Social Studies	-0.00	0.02	13829
	[1.00]	(0.670)	

Notes: The table provides summary statistics for attributes of schools, teachers (new recruits placed in upper primary only), and students collected at baseline. The first column presents means in FW schools, and the (with standard deviations in brackets); second column presents estimated differences between FW and P4P schools (with randomization inference p-values in parentheses). In Panel B, Grading Task IRT scores are standardized based on the distribution among incumbent teachers. In Panel C, student learning IRT scores are standardized based on the distribution in the experienced FW arm.

conducted via randomization inference, and are undertaken with 2,000 permutations of the experienced treatment.

4.1 Compositional margin of pay-for-performance

We study three types of compositional effects of pay-for-performance. These are impacts on: the quality of applicants; the observable skill and motivation of placed recruits on arrival; and the student learning induced by these placed recruits during their first and second year on the job.

Quality of applicants Motivated by the theoretical model sketched in Section 2.3, we begin by testing for impacts of advertised P4P on the quality of applicants to a given district-by-qualification pool (Hypothesis I). We focus on Teacher Training College final exam score since this is the only consistently measured quality-related characteristic we observe for all applicants.

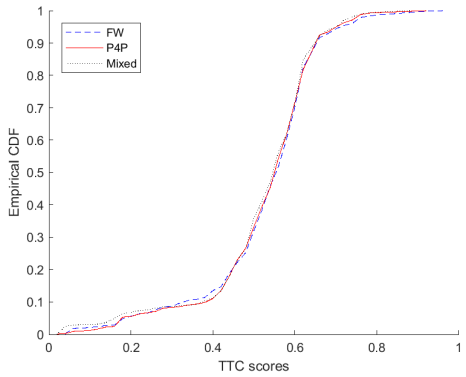
Our primary test uses a Kolmogorov-Smirnov (henceforth, KS) statistic to test the null that there is no difference in the distribution of TTC scores across advertised P4P and advertised FW labor markets. This test statistic can be written as

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

Here, $\hat{F}_{P4P}(y)$ denotes the empirical cumulative distribution function of TTC scores 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 scores 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 depicts the distribution of applicant TTC score, by advertised treatment arm. These distributions are statistically indistinguishable between advertised P4P and advertised FW. The KS test-statistic has a value of 0.026, with a p-value of 0.909. Randomization inference is well-powered, meaning that we can rule out even small effects on the TTC score distribution: a 95 percent confidence interval

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



Notes: KS test-statistic is 0.026, with a p-value of 0.909.

based on inversion of the randomization inference test rules out additive treatment effects outside of the range $[-0.027, 0.020]$. We therefore conclude that there was no meaningful impact of advertised P4P on the TTC final exam scores of applicants.[30]

Below, we move on to consider impacts of advertised P4P on the quality of applicants who were offered a post and chose to accept it—a subset that we term *placed recruits*. It is worth emphasizing that we may find results here even though there is no evidence of an impact on the distribution of TTC score of applicants. This is because, for this well-defined set of placed recruits, we have access to far richer data: lab-in-the-field instruments measuring attributes on arrival, as well as measures of student learning in the first and second years on the job.

Skill and motivation of placed recruits As Dal Bó and Finan (2016) note in a recent review, it is of interest to explore whether institutions can attract the most able into public service, as well as the most intrinsically motivated. Reflecting this, we include multidimensional skill and motivation types in the theoretical model and test the resulting hypotheses (Hypotheses II and III) using the data described in Section 3.2. Specifically, we use the Grading Task IRT score to measure a placed recruit's skill on arrival, and the framed Dictator Game share sent to capture his/her baseline intrinsic motivation.

Our primary tests use these baseline attributes of placed recruits as outcomes. For attribute x of teacher j with qualification q in district d, we estimate a regression

[30]This conclusion is further substantiated by the battery of secondary tests in Appendix C.

of the form

$$x_{jqd} = \tau_A T^A_{qd} + \gamma_q + \delta_d + e_{jqd}, \tag{2}$$

where treatment T^A_{qd} denotes the contractual condition under which a candidate applied.[31] Our test of the null hypothesis is the t statistic associated with coefficient τ_A. We obtain a randomization distribution for this t statistic under the sharp null of no effects for any hire by estimating equation (2) under the set of feasible randomizations of advertised treatments, $T^A \in \mathcal{T}^A$.

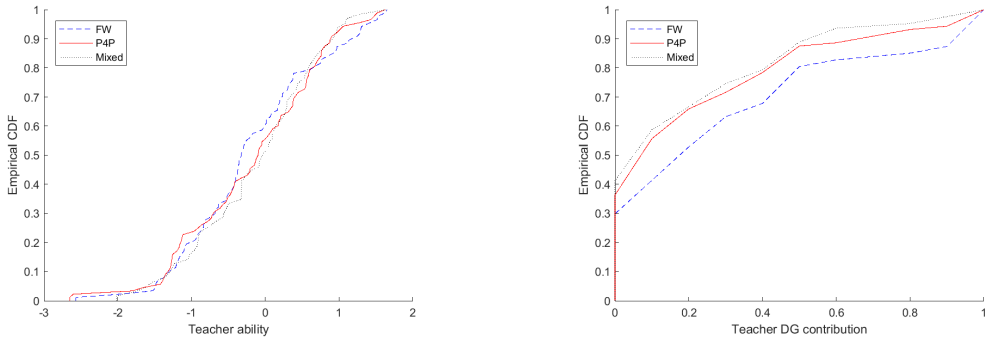
Before reporting these t statistics, it is instructive to view the data graphically. Figure 3a shows the distribution of Grading Task IRT score, and Figure 3b the framed Dictator Game share sent, by advertised treatment arm and measured on placed recruits' arrival in schools. A difference in the distributions across treatment arms is clearly visible for the measure of intrinsic motivation but not for the measure of skill. Our regression results tell the same story. In the Grading Task IRT score specification, our estimate of τ_A is -0.207, with a randomization inference p-value of 0.31. In the Dictator Game share sent specification, our estimate of τ_A is -0.108, with a randomization inference p-value is 0.02. It follows that we cannot reject the sharp null of no advertised P4P treatment effect on the measured skill of placed recruits, but we can reject the sharp null of no advertised P4P treatment effect on their measured intrinsic motivation (at the 5 percent level). Teachers recruited under advertised P4P allocated approximately 10 percentage points *less* to the students on average.

Student learning induced by placed recruits The skill and motivation of placed recruits on arrival are policy relevant insofar as these attributes translate into effectiveness in the classroom. To assess this, we combine experimental variation in the advertised contracts to which these recruits applied, with the second-stage randomization in experienced contracts under which they worked. This allows us to estimate the impact of advertised P4P on the student learning induced by these recruits, holding constant the experienced contract—a pure compositional effect (Hypothesis IV).

Our primary test is derived from estimates on student-subject-year level data.

[31] 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 were placed) under the pure FW treatment are considered as the omitted category here, to which P4P recruits will be compared.

Figure 3: Distribution of placed recruit attributes on arrival, by advertised treatment arm



(a) Grading task score

(b) Dictator Game contribution

Notes: In Figure 3a, the t statistic for a difference in mean Grading Task IRT score across the P4P and FW treatments is -0.207, with a p-value of 0.31. In Figure 3b, the t statistic for a difference in mean DG share sent across the P4P and FW treatments is -0.108, with a p-value of 0.02.

The advertised treatment about which a given student's performance is informative depends on the identity of the placed recruit teaching that particular subject via qualification type and district. We denote this by T^A_{qd} for teacher j 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.

We pool data across the two years of intervention to estimate a specification of the type

$$z_{ibksr} = \tau_A T^A_{qd} + \tau_E T^E_s + \lambda_I I_j + \lambda_E T^E_s I_j + \rho_{br} \bar{z}_{ks,r-1} + \delta_d + \psi_r + e_{ibksr} \qquad (3)$$

for the learning outcome of student i in subject b, stream k, school s, and round r. We define $j = j(bksr)$ as an identifier for the teacher assigned to that subject-stream-school-round. The variable I_j is an indicator for whether the teacher is an incumbent, and the index $q = q(j)$ denotes the qualification type of teacher j 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 placed in that stream, and its coefficient, ρ_{br}, is subject- and round-specific. The coefficient of interest is τ_A.

The theoretical model of Appendix B, as well as empirical evidence from other contractual settings (Einav et al., 2013), suggests that pay-for-performance may induce selection on the *responsiveness* to performance incentives. If so, then the impact of advertised treatment will depend on the contractual environment into which recruits are placed. Consequently, we also estimate a specification that allows advertised treatment effects to differ by experienced treatment, including an interaction term between the two treatments. This interacted model takes the form

$$z_{ibksr} = \tau_A T^A_{qd} + \tau_E T^E_s + \tau_{AE} T^A_{qd} T^E_s + \lambda_I I_j + \lambda_E T^E_s I_j + \rho_{bgr} \bar{z}_{ks,r-1} + \delta_d + \psi_r + e_{ibksr}. \quad (4)$$

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).

We estimate equations (3) and (4) by a linear mixed effects model, allowing for normally distributed random effects at the student-round level.[32] Randomization inference is used throughout. To do so, we focus on the distribution of the estimated z-statistic (i.e., the coefficient divided by its estimated standard error), which allows rejections of the sharp null of no effect on any student's performance to be interpreted, asymptotically, as rejection of the non-sharp null that the coefficient is equal to zero (DiCiccio and Romano, 2017). Inference for τ_A is undertaken by permutation of the advertised treatment, $T^A \in \mathcal{T}^{\mathcal{A}}$, while inference for τ_E likewise proceeds by permuting the experienced treatment $T^E \in \mathcal{T}^{\mathcal{E}}$. To conduct inference about the interaction term, τ_{AE} in equation (4), we simultaneously permute both dimensions of the treatment, considering pairs (T^A, T^E) from the set $\mathcal{T}^{\mathcal{A}} \times \mathcal{T}^{\mathcal{E}}$.

Results are presented in Table 3. Pooling across years, the compositional effect of advertised P4P is small in point-estimate terms, and statistically indistinguishable from zero (Model A, first row). We do not find evidence of selection on responsiveness to incentives; if anything, the effect of P4P is stronger among recruits who applied under advertised FW contracts, although that difference is not statistically

[32] In our pre-analysis plan, simulations using the blinded data indicated that the linear mixed effects model with a student-round normal random effects would maximise statistical power. We found precisely this in the unblinded data. For completeness, and purely as supplementary analysis, we also present estimates and hypotheses tests via ordinary least squares. See Appendix Table A.3. These OLS estimates are generally larger in magnitude and stronger in statistical significance.

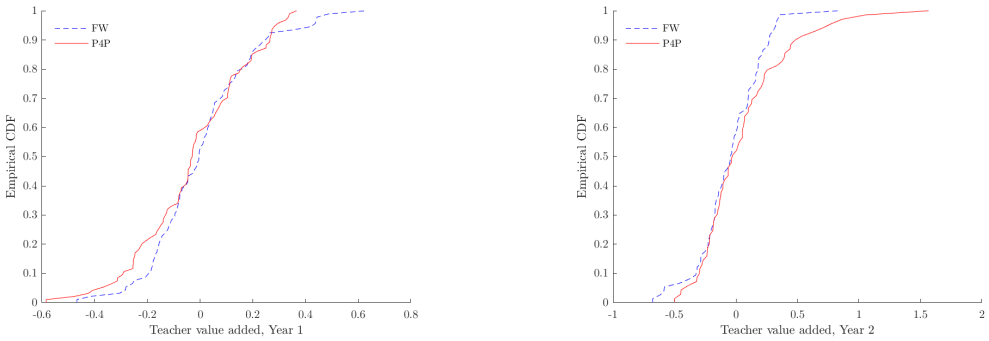
Table 3: Impacts on student learning, linear mixed effects model

	Pooled	Year 1	Year 2
Model A: Direct effects only			
Advertised P4P	0.01	-0.03	0.05
	[-0.05, 0.11]	[-0.06, 0.08]	[-0.06, 0.19]
	(0.70)	(0.16)	(0.18)
Experienced P4P	0.11	0.06	0.16
	[-0.00, 0.22]	[-0.06, 0.15]	[0.03, 0.30]
	(0.03)	(0.17)	(0.01)
Experienced P4P × Incumbent	-0.07	-0.05	-0.09
	[-0.23, 0.11]	[-0.22, 0.12]	[-0.29, 0.09]
	(0.35)	(0.54)	(0.26)
Model B: Interactions between advertised and experienced contracts			
Advertised P4P	0.03	-0.01	0.06
	[-0.07, 0.16]	[-0.07, 0.09]	[-0.08, 0.23]
	(0.47)	(0.63)	(0.21)
Experienced P4P	0.14	0.08	0.19
	[0.02, 0.25]	[-0.04, 0.19]	[0.05, 0.34]
	(0.01)	(0.13)	(0.01)
Advertised P4P × Experienced P4P	-0.04	-0.03	-0.03
	[-0.18, 0.11]	[-0.19, 0.11]	[-0.24, 0.17]
	(0.53)	(0.64)	(0.69)
Experienced P4P × Incumbent	-0.09	-0.07	-0.11
	[-0.36, 0.20]	[-0.37, 0.23]	[-0.49, 0.18]
	(0.42)	(0.56)	(0.36)
Observations	154594	70821	83773

Notes: For each estimated parameter, or combination of parameters, the table reports the point estimate (stated in standard deviations of student learning), 95 percent confidence interval in brackets, and p-value in parentheses. Randomization inference is conducted on the associated z statistic. The measure of student learning is based on the empirical Bayes estimate of student ability from a two-parameter IRT model, as described in Section 3.3.

significant and the 95 percent confidence interval for this estimate is wide (Model B, third row). The effect of advertised P4P on student learning does, however, appear to strengthen over time. By their second year on the job, recruits who applied under advertised P4P delivered 0.05 standard deviations of additional learning per year on average compared to their FW applicant counterparts. OLS estimates of this effect are even larger, at 0.09 standard deviations, and significant at the 10 percent level, as shown in Table A.3.

Figure 4: Teacher value added among recruits, by advertised treatment and year



(a) Year 1 (b) Year 2

Notes: The figures plot distributions of teacher value added under advertised P4P and advertised FW in Years 1 and 2. Value-added models estimated with school fixed effects. Randomization inference p-value for equality in distributions between P4P and FW applicants, based on one-sided KS test, is 0.796 using Year 1 data; 0.123 using Year 2 data; and 0.097 using pooled estimates of teacher value added (not pre-specified).

For the purposes of interpretation, it is useful to recast the data in terms of teacher value added. As detailed in Appendix D, we do so by estimating a teacher valued-added (TVA) model that controls for students' lagged test scores, as well as school fixed effects, with the latter absorbing differences across schools attributable to the experienced P4P treatment. This TVA model gives a sense of magnitude to the student learning estimates in Table 3. Applying the Year 2 point estimate for the effect of advertised P4P would raise a teacher from the 50th to above the 76th percentile in the distribution of (empirical Bayes estimates of) teacher value added for placed recruits who applied under FW. The TVA model also reveals the impact of advertised P4P on the distribution of teacher effectiveness. Figure 4b

shows that the distribution of teacher value added among recruits in their second year on the job is better, by first order stochastic dominance, under advertised P4P than advertised FW. This finding is consistent with the view that a contract that rewards the top quintile of teachers attracts individuals who deliver greater learning.

4.2 Effort margin of pay-for-performance

Having studied the type of individuals applying to, and being placed in, upper-primary posts, we now consider the activities undertaken by these new recruits.

Student learning induced by placed recruits We start by using the two-tiered experimental variation to estimate the impact of experienced P4P on the student learning induced by the placed recruits, holding constant the advertised contract—a pure effort effect (Hypothesis V). Our primary test uses the specification in equation (3), again estimated by a linear mixed effects model. The coefficient of interest is now τ_E. To investigate possible 'surprise effects' from the re-randomization, we also consider the interacted specification of equation (4). In this model, τ_E gives the 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 effect of experienced P4P among recruits who applied under P4P contractual conditions (a comparison of groups b and d). If recruits are disappointed, because it is groups b and c who received the surprise, τ_E should be smaller than $\tau_E + \tau_{AE}$.

Results are presented in Table 3. Pooling across years, the effect of experienced P4P is 0.11 standard deviations of additional learning per year (Model A, second row). The randomization inference p-value is 0.03, implying that we can reject the sharp null of no experienced P4P treatment effect on placed recruits at the 5 percent level. We do not find evidence of disappointment caused by the re-randomization. The interaction term is insignificant (Model B, third row) and, in point-estimate terms, τ_E is larger than $\tau_E + \tau_{AE}$. As was the case for the compositional margin, the effort effect of experienced P4P on student learning appears to strengthen over time. By their second year on the job, new recruits working under P4P contracts delivered 0.16 standard deviations of additional learning per year on average compared to their FW counterparts.

Dimensions of the composite performance metric The results in Table 3 speak to the obvious policy question, namely whether there are impacts of advertised

and experienced P4P contracts on student learning, measured as empirical Bayes estimates from a two-parameter student-level IRT model. For completeness, and to gain an understanding into mechanisms, we complete our analysis by studying whether there are impacts on the *contracted* metrics which are calculated at teacher-level (Hypothesis VI). For these tests, we use the following specifications:

$$m_{jqsdr} = \tau_A T^A_{qd} + \tau_E T^E_s + \lambda_I I_j + \lambda_E T^E_s I_j + \gamma_q + \delta_d + \psi_r + e_{jqsdr} \quad (5)$$

$$m_{jqsdr} = \tau_A T^A_{qd} + \tau_E T^E_s + \tau_{AE} T^A_{qd} T^E_s + \lambda_I I_j + \lambda_E T^E_s I_j + \gamma_q + \delta_d + \psi_r + e_{jqsdr}, \quad (6)$$

for the metric of teacher j with qualification q in school s of district d, as observed in post-treatment round r. As above, the variable I_j is an indicator for whether the teacher is an incumbent (recall that T^A_{qd} is always zero for incumbents).[33] 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 stated in our pre-analysis plan, we therefore estimate equations (5) and (6) with a round-school random-effects estimator to improve efficiency. The permutations of treatments used for inferential purposes mirror those above.

Results are reported in Table 4 and, to the extent available, are based on pooled data.[34] Consistent with the pooled results in Table 3, we see a positive and significant impact of experienced P4P on both the summary metric and the learning sub-component. The specifications with teacher inputs as dependent variables suggest that this impact on student learning is driven, at least in part, by improvements in teacher presence and pedagogy. Teacher presence was 8 percentage points higher among recruits who experienced the P4P contract compared to recruits who experienced the FW contract; an impact that is statistically significant at the 1 percent level and sizeable in economic terms given that baseline teacher presence was already 90 percent. Recruits who experienced P4P were more effective in their classroom practices than recruits who received FW by 0.10 points, although this impact is weaker in terms of statistical significance.

[33] Note that any attribute of recruits themselves, even if observed at baseline, suffers from the 'bad controls' problem, as the observed values of this covariate could be an outcome of the advertised treatment. These variables are therefore not included as independent variables.

[34] As discussed in Section 3.4, FW schools only received unannounced visits to measure teacher inputs in Year 2.

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

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

Notes: For each estimated parameter, the table reports the point estimate, 95 percent confidence interval in brackets, and p-value (or for FW means, standard deviations) in parentheses. Randomization inference is conducted on the associated t statistic. All estimates are pooled across years, but outcomes are observed in the FW arm during the second year only. Outcomes are constructed at teacher-round-level as follows: *preparation* is a binary indicator for existence of a lesson plan on a randomly chosen spot-check day; *presence* is the fraction of spot-check days present at the start of the school day; *pedagogy* is the classroom observation score, measured on a four-point scale; and *pupil learning* is the Barlevy-Neal percentile rank. The *summary metric* places 50 percent weight on learning and 50 percent on teacher inputs, and is measured in percentile ranks.

4.3 Dynamic effects

Our two-tiered 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 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.[35]

Retention effects We begin by exploring whether experienced P4P affects retention rates among recruits. Specifically, we look for an impact 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 pay-for-performance in Year 1, although before the performance awards were announced. To do so, we use a linear probability model of the form

$$\Pr[employed_{iqd2} = 1] = \tau_E T_s^E + \gamma_q + \delta_d, \tag{7}$$

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 δ_d are the usual subject-family qualification and district indicators.

[35] We emphasise that this material is 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.

As the first column of Table 5 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 5: Retention of placed 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...		Grading Task	Dictator Game
Observations	249	238	238

Notes: For each estimated parameter, the table reports the point estimate and p-value in brackets. Randomization inference is conducted on the associated t statistic. In each column the outcome is a binary indicator taking the value of 1 if the teacher is still employed at the start of Year 2. In the second column, the specification includes an interaction of experienced treatment with the teacher's baseline Grading Task IRT score; in the third column, the interaction is with the teacher's share sent in the baseline framed Dictator Game. 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 (5) and (6). As the second column of Table A.4 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.

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. To explore this, we test whether experienced P4P induces differentially skilled recruits to be retained. Here, we use a teacher's performance on the baseline Grading Task in the primary subject he/she teaches to obtain an IRT estimate of his/her ability in this subject, denoted z_i, and

estimate an interacted model of the form

$$\Pr[employed_{iqd2} = 1] = \tau_E T_s^E + \zeta T_s^E z_i + \beta z_i + \gamma_q + \delta_d. \tag{8}$$

Inference for the key parameter, ζ, is undertaken by performing randomization inference for alternative assignments of the school-level experienced treatment indicator. As the second column of Table 5 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. Hence, there is no evidence that experienced P4P induces differentially skilled recruits to be retained.

We also test whether experienced P4P induces differentially intrinsically motivated recruits to be retained. Here, we use the contribution sent in the framed Dictator Game played by all recruits at baseline, denoted x_i, and re-estimate the interacted model in equation (8), replacing z_i with x_i. As the third column of Table 5 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. Hence, there is also no evidence that experienced P4P induces differentially intrinsically motivated recruits to be retained.

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

$$y_{isd2} = \tau_E T_s^E + \rho y_{isd0} + \gamma_q + \delta_d + e_{isd}, \tag{9}$$

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). As the first column of Table 6 reports, our estimate of τ_E in the Grading Task specification is 0.57, with a randomization inference p-value of 0.63. Our estimate of τ_E in the Dictator Game specification is -0.04, with a randomization inference p-value of 0.06. Both estimates are small in magnitude and, in the case of the Dictator Game share sent, we reject the sharp null only at the 10 percent level. Hence, to the extent that contributions in the Dictator Game are positively associated with teachers' intrinsic motivation, we find no evidence that

Table 6: Characteristics of retained recruits at endline

	Grading Task	Dictator Game
Experienced P4P	0.57	-0.04
	[0.63]	[0.06]
Observations	170	169

Notes: For each estimated parameter, the table reports the point estimate and p-value in brackets. Randomization inference is conducted on the associated t statistic. In the first column, the outcome is the Grading Task score of the teacher at endline (measured on (raw) a scale from 0 to 30); and in the second column, the outcome is the teacher's share sent in the framed Dictator Game played at endline. All specifications include the outcome measured at baseline and controls for district and subject-of-qualification.

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.[36]

Before moving on, it is worth noting that the Dictator Game result could be interpreted as weak evidence that the experience of P4P contracts crowded out the intrinsic motivation of recruits. We do not have any related measures observed at both baseline and endline with which to further probe *changes* in motivation. However, we do have a range of related measures in Year 2: job satisfaction, likelihood of leaving, and positive/negative affect.[37] As Table A.4 shows, there is no statistically significant impact of experienced P4P on any of these measures. Although the confidence intervals reflect the (presumably) noisy nature of the survey responses, we can rule out economically meaningful negative impacts of experienced P4P on these measures of motivation.

Further substantiating this point, Table A.5 shows the distribution of answers to the endline survey question: "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?"[38] Across all recruits, the propor-

[36] Although repeated play of lab experimental games may create interpretation concerns in some contexts, there are several factors that allay this concern here. First, unlike strategic games, the so-called 'Dictator Game' has no second 'player' about whose behavior the dictator can learn. Second, the two rounds of play were fully two years apart.

[37] 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 to capture the overall attitude of teachers. These measures were constructed using data from the endline teacher survey.

[38] We follow the phrasing used in the surveys conducted by Muralidharan and Sundararaman

tion giving a favourable answer is high, and never lower than 75 percent. In terms of Figure 1, it was group a, recruits who both applied for and experienced FW, who had the most negative view of pay-for-performance, and group c, recruits who applied for FW but experienced P4P, who had the most positive view. Hence it seems that it was the idea, rather than the reality, of pay-for-performance that was unpopular with (a minority of) recruits.[39]

5 Discussion

Compositional margin To recap from Section 4.1, we find no evidence of an advertised treatment impact on the type of individuals who *apply* for upper-primary teaching posts in study districts, but we do find evidence of an advertised treatment impact on the type of individuals who are *placed* into study schools. We draw the following conclusions from these results.

Potential applicants were aware of, and responded to, the labor market intervention. The difference in distributions across advertised treatment arms in Figure 3b (Dictator Game share sent) and Figure 4b (teacher valued added in Year 2) show that the intervention changed behavior. Since these differences are for placed recruits not applicants, it is possible that this behavior change was on the labor demand rather than supply-side. In Appendix C, we show that there is no evidence of an advertised treatment impact on hiring patterns, and hence conclude that these differences do indeed reflect a supply-side response.

This supply-side response was, if anything, beneficial for student learning. The P4P contract negatively selected-in the attribute measured by the baseline Dictator Game. However, Appendix Table D.1 shows that the rank correlation between the baseline DG share sent by recruits and their teacher value added is small and not statistically significant. Consistent with this, our primary test rules out meaningful negative effects of advertised P4P on student learning. In fact, our supplementary analyses—the OLS estimates in Appendix Table A.3 and the distributions of teacher value added in Figure 4—point to *positive* effects on learning by recruits' second year on the job. It therefore appears that only positively selected attribute(s) mattered, at least in the five core subjects that we assessed.

(2011a). Answers are on a 5-point Likert scale ranging from 'very unfavourable' to 'very favourable'.

[39] Consistent with our failure to find 'surprise effects' in student learning, there is no evidence that the re-randomization resulted in hostility toward pay-for-performance; if anything the reverse.

Districts would struggle to achieve this compositional effect directly via the hiring process. The positively selected attribute(s) were not evident in the metrics observed at baseline—either in TTC scores which districts already use, or in the Grading Task scores that they could in principle adopt.[40] This suggests that there is not an obvious demand-side policy alternative to contractually induced supply-side selection.

Effort margin To recap from Section 4.2, we find evidence of a positive impact of experienced P4P on student learning, which is considerably larger (almost tripling in magnitude) in recruits' second year on the job. In light of Section 4.3, we draw the following conclusions from these results.

The additional learning achieved by recruits working under P4P, relative to recruits working under FW, is unlikely to be due to selection-out—the compositional margin famously highlighted by Lazear (2000). Within-year teacher turnover was limited by design. Between-year turnover did happen but cannot explain the experienced P4P effect. In Appendix D, we show that the rank correlation between recruits' baseline Grading Task IRT score and their teacher value added is positive. However, in Section 4.3 we reported that, if anything, selection-out on baseline teacher skill runs the wrong way to explain the experienced P4P effect.

Neither is the experienced P4P effect likely to be due to within-teacher changes in skill or motivation. We find no evidence that recruits working under P4P made greater gains on the Grading Task from baseline to endline than did recruits working under FW. As already noted, recruits' Dictator Game share sent is not a good predictor of teacher value added. But even if it were, we find no evidence that recruits working under P4P contributed more from baseline to endline than did recruits working under FW, if anything the reverse.

Instead, the experienced P4P effect is most plausibly driven by teacher effort. This conclusion follows from the arguments above and the direct evidence that recruits working under P4P provided greater inputs than did recruits working under FW. Specifically, the P4P contract encouraged recruits to be present in school more often and to use better pedagogy in the classroom, behaviors that were incentivized components of the 4P performance metric.

[40] An alternative explanation for the null KS test on applicant TTC scores is that individuals applied everywhere. If this were true, we would expect to see most candidates make multiple applications, and a rejection of the null in a KS test on *placed recruits'* TTC scores (if the supply-side response occurred at acceptance rather than application). We do not see either in the data.

Total effect The total effect of the P4P contract combines both the advertised and experienced impacts: $\tau_A + \tau_E$. Our estimate for Year 2 is $0.05 + 0.16 = 0.21$ standard deviations of additional learning per year, which is statistically significant at the one percent level. The discussion in this section suggests that roughly one quarter can be attributed to supply-side selection at the recruitment stage, with the remaining three quarters arising from increased teacher effort. At a minimum, our results suggest that fears that crowd-out in the supply of effective public-sector employees might dominate any effort-margin responses to pay-for-performance contracts appear overstated.

An interesting question is why this effect is so much stronger in Year 2 compared to Year 1, particularly on the effort margin. One interpretation is that this is because it takes time for recruits to settle in to the job and for the signal to noise ratio in our student learning measures to improve (Staiger and Rockoff, 2010). Consistent with this interpretation, we note that the impact of experienced P4P on incumbents did not increase in the second year. This interpretation suggests that Year 2 effects are the best available estimate of longer-term impacts.

6 Conclusion

This two-tier, two-year, randomized controlled trial featuring extensive data on teachers—their skills and motivations before starting work, multiple dimensions of their on-the-job performance, and whether they left their jobs—offers new insights into the compositional and effort margins of pay-for-performance. We found that potential applicants were aware of, and responded to, the first-tier labor market intervention. This supply-side response to advertised P4P was, if anything, beneficial for student learning. We also found a positive impact of experienced P4P that appears to stem from increased teacher effort, rather than selection-out or changes in measured skill or intrinsic motivation. By the second year of the study, we estimate the total effect of P4P to be 0.21 standard deviations of pupil learning per year. 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 pedagogy.

Our experiment was conducted at near national scale in Rwanda, covering the bulk of teacher recruitment for 2016. We used three percent of teacher salaries as the expected value of the bonus to ensure that the payment budget would not be out

of reach for any implementing government, in Rwanda or elsewhere.[41] We worked closely with the Government to construct a multidimensional P4P metric that would not narrowly emphasize any single aspect of teacher performance. Moreover, when measuring learning, we employed the Barlevy and Neal (2012) pay-for-percentile approach that aims to give all teachers a fair chance, regardless of the composition of the students they teach. Our study was therefore intended to evaluate the effects of a policy that would be reasonable on its face and feasible at scale.

There are nonetheless limitations of our work. Inasmuch as the impacts on either the compositional or effort margin might be different after five or ten years, there is certainly scope for further study of this topic in low- and middle-income countries. An often-discussed limitation of pay-for-performance is the challenge associated with measurement. In terms of pupil learning, the minimum requirement for the P4P contract we study is a system of repeated annual assessments across grades and key subjects.[42] Measurement of the other aspects of performance—teacher presence, preparation, and pedagogy—is less complex and can in principle be conducted by head teachers (or other school or district staff) at relatively modest cost.

Rwanda's labor market has a characteristic unusual for low- and middle-income countries: it has no public sector pay premium, and consequently many of those qualified to teach choose not to, making it more similar to high- income country labor markets in this regard. Whether the multidimensional, pay-for-percentile type of contract that we found to be effective in Rwanda—improving performance without dampening employee satisfaction—might do the same in high-income countries, remains an open question, for the education sector and beyond.

[41] The three percent figure is broadly in line with annual increments, and with discretionary pay in other sectors under Rwanda's *imihigo* system of performance contracts for civil servants.

[42] Such a system does not yet exist in Rwanda but may soon be introduced, as part of the recently announced 'comprehensive assessment' program. See, e.g., https://www.newtimes.co.rw/opinions/mineducs-new-guide-student-assessment-triggers-debate.

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.

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.

_ , _ , Edward Davenport, and Scott S. Lee, "Losing prosociality in the quest for talent? Sorting, selection, and productivity in the delivery of public services," *American Economic Review*, forthcoming.

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

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

Bau, Natalie and Jishnu Das, "Teacher value-added in a low-income country," *American Economic Journal: Economic Policy*, forthcoming.

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.

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.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers, "Missing in Action: Teacher and Health Worker Absence in Developing Countries," *Journal of Economic Perspectives*, Winter 2006, *20* (1), 91–116.

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.

Chingos, Matthew M and Martin R West, "Do more effective teachers earn more outside the classroom?," *Education Finance and Policy*, 2012, *7* (1), 8–43.

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.

Deci, Edward L. and Richard M. Ryan, *Intrinsic motivation and self-determination in human behavior*, New York: Plenum, 1985.

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.

DiCiccio, Cyrus J and Joseph P Romano, "Robust permutation tests for correlation and regression coefficients," *Journal of the American Statistical Association*, 2017, *112* (519), 1211–1220.

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

Einav, Liran, Amy Finkelstein, Stephen P. Ryan, Paul Schrimpf, and Mark R. Cullen, "Selection on moral hazard in health insurance," *American Economic Review*, 2013, *103* (1), 178–219.

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

Finan, Frederico, Benjamin A Olken, and Rohini Pande, "The personnel economics of the state," in "Handbook of Field Experiments," Vol. 2, Elsevier, 2017, pp. 467–514.

Gilligan, Daniel O., Naureen Karachiwalla, Ibrahim Kasirye, Adrienne Lucas, and Derek A. Neal, "Educator incentives and educational triage in rural primary schools," *Journal of Human Resources*, forthcoming.

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.

Imbens, Guido W and Donald B Rubin, *Causal inference for statistics, social, and biomedical sciences: An introduction*, Cambridge, U.K.: Cambridge University Press, 2015.

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

Jackson, C. Kirabo, Jonah E. Rockoff, and Douglass O. Staiger, "Teacher effects and teacher-related policies," *Annual Review of Economics*, 2014, *6*, 801–825.

Kane, Thomas J and Douglas O Staiger, "Estimating teacher impacts on student achievement: An experimental evaluation," NBER Working Paper 14607 December 2008.

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

Krepps, David, "Intrinsic motivation and extrinsic incentives," *American Economic Review*, 1997, *87* (2), 359–64.

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.

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, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani, "Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania," *Quarterly Journal of Economics*, 2019, *134* (3), 1627–1673.

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.

Neal, Derek A, "The design of performance pay in education," in Eric A. Hanushek, Stephen J. Machin, and Ludger Woessmann, eds., *Handbook of the Economics of Education*, Vol. 4, Amsterdam: North Holland, 2011.

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

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

Staiger, Douglas O and Jonah E Rockoff, "Searching for effective teachers with imperfect information," *Journal of Economic Perspectives*, 2010, *24* (3), 97–118.

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.

Stecher, Brian M., Deborah J. Holtzman, Michael S. Garet, Laura S. Hamilton, John Engberg, Elizabeth D. Steiner, Abby Robyn, Matthew D. Baird, Italo A. Gutierrez, Evan D. Peet, Iliana Brodziak de los Reyes, Kaitlin Fronberg, Gabriel Weinberger, Gerald P. Hunter, and Jay Chambers, "Improving teacher effectiveness: Final Report: The intensive partnerships for effective teaching through 2015-2016," Santa Monica, CA: RAND Corporation 2018.

For online publication

Appendix A Supplemental figures and tables

Figure A.1: Study profile

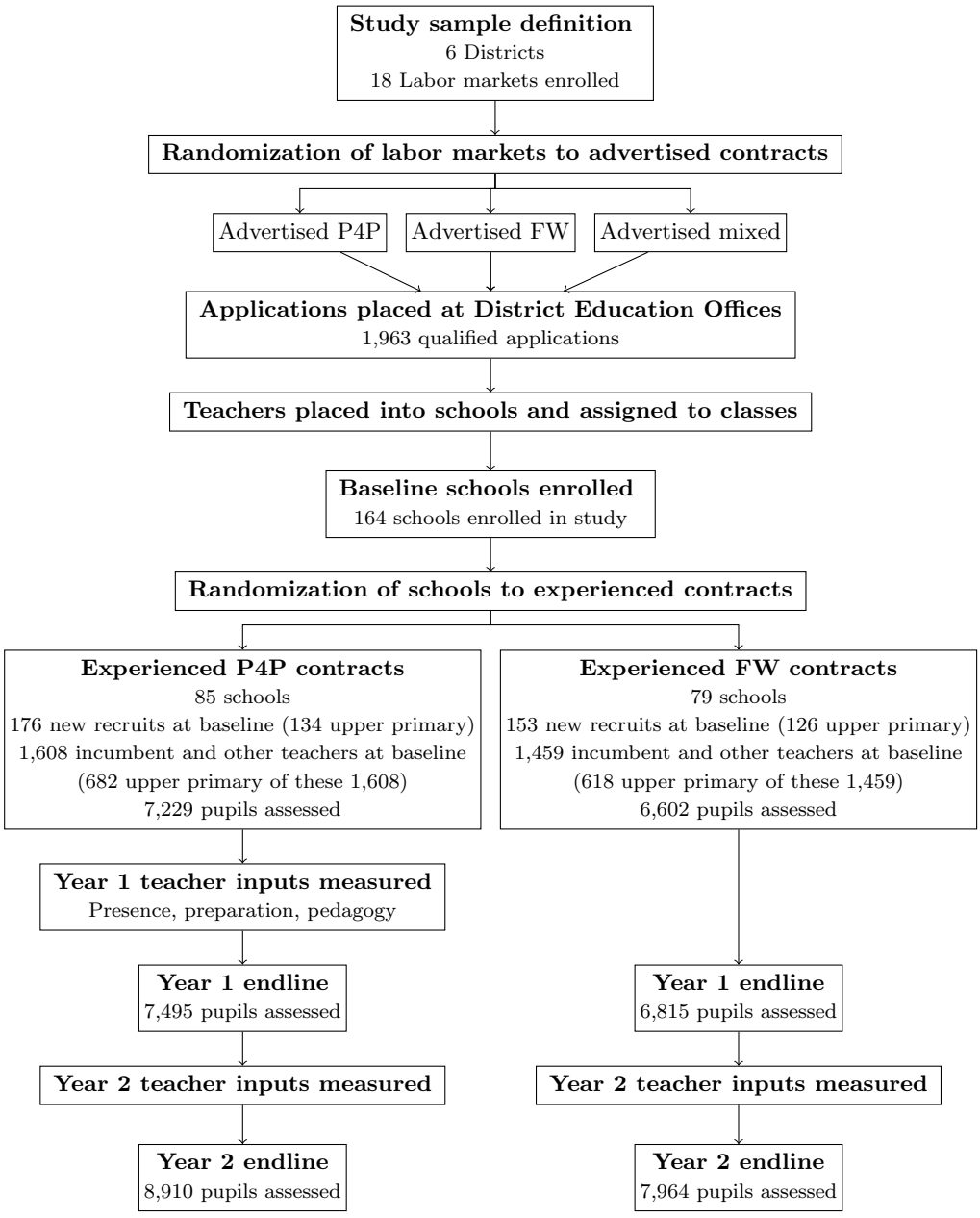


Table A.1: Summary of hypotheses, outcomes, samples, and specifications

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. (10)	$\mathcal{T}^A$
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. (10)	$\mathcal{T}^A$
TTC exam scores	Universe of application, weighted by probability of placement	t_A in eq. (10)	$\mathcal{T}^A$
Number of applicants	Universe of applications	t_A in eq. (11)	$\mathcal{T}^A$
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. (2)	$\mathcal{T}^A$
HYPOTHESIS III: ADVERTISED P4P INDUCES DIFFERENTIALLY 'INTRINSICALLY' MOTIVATED RECRUITS TO BE PLACED IN SCHOOLS			
*Dictator-game donations	Placed recruits	t_A in eq. (2)	$\mathcal{T}^A$
Perry PSM instrument	Placed recruits retained through Year 2	t_A in eq. (2)	$\mathcal{T}^A$
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. (3)	$\mathcal{T}^A$
Student assessments	Pooled Year 1 & Year 2 students	t_A and t_{A+AE}; t_{AE} in eq. (4)	$\mathcal{T}^A$ $\mathcal{T}^A \times \mathcal{T}^E$

Continues...

Table A.1, continued

Outcome	Sample	Test statistic	Randomization inference
Student assessments	Year 1 students	t_A in eq. (3)	$\mathcal{T}^A$
Student assessments	Year 2 students	t_A in eq. (3)	$\mathcal{T}^A$
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. (3)	$\mathcal{T}^E$
Student assessments	Pooled Year 1 & Year 2 students	t_E and t_{E+AE}; t_{AE} in eq. (4)	$\mathcal{T}^E$ $\mathcal{T}^A \times \mathcal{T}^E$
Student assessments	Year 1 students	t_E in eq. (3)	$\mathcal{T}^E$
Student assessments	Year 2 students	t_E in eq. (3)	$\mathcal{T}^E$
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. (5)	$\mathcal{T}^A$
Composite 4P metric	Teachers, pooled Year 1 (experienced P4P only) & Year 2	t_A and t_{A+AE}; t_E and t_{E+AE}; t_{AE} in eq. (6)	$\mathcal{T}^A$ $\mathcal{T}^E$ $\mathcal{T}^A \times \mathcal{T}^E$
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 (3) and (4), 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}.

Table A.2: 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 for upper-primary teachers only. Overall score for the classroom observation is the average of four components: lesson objective, teaching activities, use of assessment, and student engagement, with each component scored on a scale from 0 to 3.

Table A.3: Impacts on student learning, OLS model

	Pooled	Year 1	Year 2
Model A: Direct effects only			
Advertised P4P	0.03	-0.03	0.09
	[-0.03, 0.17]	[-0.11, 0.11]	[-0.02, 0.25]
	(0.29)	(0.49)	(0.06)
Experienced P4P	0.13	0.10	0.17
	[0.00, 0.27]	[-0.04, 0.22]	[0.01, 0.33]
	(0.01)	(0.06)	(0.02)
Experienced P4P × Incumbent	-0.09	-0.10	-0.09
	[-0.37, 0.19]	[-0.39, 0.20]	[-0.48, 0.20]
	(0.42)	(0.39)	(0.46)
Model B: Interactions between advertised and experienced contracts			
Advertised P4P	0.04	-0.03	0.12
	[-0.07, 0.24]	[-0.15, 0.17]	[-0.03, 0.36]
	(0.41)	(0.58)	(0.08)
Experienced P4P	0.14	0.10	0.17
	[0.01, 0.27]	[-0.03, 0.22]	[-0.01, 0.36]
	(0.01)	(0.11)	(0.03)
Advertised P4P × Experienced P4P	-0.02	0.01	-0.05
	[-0.24, 0.21]	[-0.19, 0.24]	[-0.37, 0.21]
	(0.75)	(0.98)	(0.66)
Experienced P4P × Incumbent	-0.09	-0.09	-0.09
	[-0.69, 0.56]	[-0.66, 0.53]	[-0.86, 0.53]
	(0.65)	(0.61)	(0.67)
Observations	154594	70821	83773

Notes: For each estimated parameter, or combination of parameters, the table reports the point estimate (stated in standard deviations of student learning), 95 percent confidence interval in brackets, and p-value in parentheses. Randomization inference is conducted on the associated t statistic. The measure of student learning is based on the empirical Bayes estimate of student ability from a two-parameter IRT model, as described in Section 3.3.

Table A.4: 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: For each estimated parameter, or combination of parameters, the table reports the point estimate (stated in standard deviations of student learning), 95 percent confidence interval in brackets, and p-value in parentheses. Randomization inference is conducted on the associated t statistic. Outcomes are constructed as follows: *job satisfaction* is scored on a 7-point scale with higher numbers representing greater satisfaction; *likehood of leaving* is a binary indicator coded to 1 if the teacher responds that they are likely or very likely to leave their job at the current school over the coming year; *positive affect* and *negative affect* are standardized indices derived from responses on a 5-point Likert scale.

Table A.5: Teacher attitudes toward pay-for-performance 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.

Appendix B 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. 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'.[43]

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 $U\left[\underline{\varepsilon}, \overline{\varepsilon}\right]$, and are independent across employments.

Compensation schemes 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. In the advertised FW treatment, individuals choose between: (i) an education contract

[43] 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 pay-for-performance contract, and (iii) declining and applying for a job in an outside sector on a different performance 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.

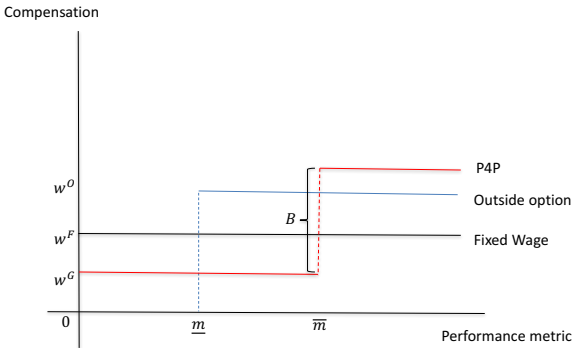
1. Outside options and education contract offers are announced.
2. Nature chooses type (τ, θ).
3. Individuals observe their type (τ, θ), and choose which sector to appply to.
4. Employers hire (at random) from the set of applicants.
5. *Surprise* re-randomization occurs.
6. Individuals make effort choice 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 B.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.

Figure B.1: Compensation schemes in the numerical example



Effort incentives Effort choices under the three compensation schemes are:

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

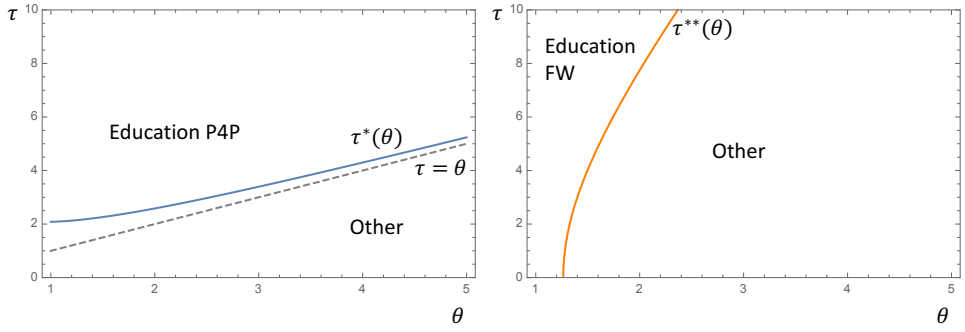
where we have used the fact that ε is drawn from a uniform distribution. Intuitively, effort incentives are higher under P4P than under FW, i.e. $e^P > e^F$.

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 B.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

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



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.

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.

Appendix C Teacher hiring process

Secondary analysis of applications

Our pre-analysis plan included a small number of secondary tests of Hypothesis I (see Table A.1). Three of these tests use estimates from TTC score regressions of the form

$$y_{iqd} = \tau_A T^A_{qd} + \gamma_q + \delta_d + e_{iqd}, \quad \text{with weights } w_{iqd}, \tag{10}$$

where y_{iqd} denotes the TTC exam score of applicant teacher i with qualification q in district d and treatment T^A_{qd} denotes the contractual condition under which a candidate applied. The weighted regression parameter τ_A estimates the difference in (weighted) mean applicant skill induced by advertised P4P. The fourth test is for a difference in the number of applicants by treatment status, conditional on district and subject-family indicators. Here, we use a specification of the form

$$\log N_{qd} = \tau_A T^A_{qd} + \gamma_q + \delta_d + e_{qd}, \tag{11}$$

where q indexes subject families and d indexes districts; N_{qd} measures the number of qualified applicants in each district.[44]

To undertake inference about these differences 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.[45] 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 $\tau_A = 0$.

Results are in Table C.1. The first column restates the confidence interval and p-value from the KS test for comparison purposes. The second column reports results for the TTC score regression where all observations are weighted equally (i.e. a random hiring rule, as assumed in the theory). Our estimate of τ_A is -0.001. The

[44] '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.

[45] 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).

Table C.1: Secondary tests of impacts on teacher ability in application pool

	KS	Unweighted	Empirical weights	Top	Number of Applicants
Advertised P4P	n.a. [-0.027, 0.020] (0.909)	-0.001 [-0.044, 0.043] (0.984)	-0.001 [-0.040, 0.039] (0.948)	-0.009 [-0.026, 0.009] (0.331)	-0.040 [-0.360, 0.386] (0.811)
Observations	1715	1715	1715	1715	18

Notes: The first column shows the confidence interval in brackets, and the p-value in parentheses, from the primary KS test discussed in Section 4.1. ('n.a' reflects the lack of a corresponding point estimate.) The second through fourth columns report the point estimate of τ_A from the applicant TTC exam score specification in (10) with the stated weights. The fifth column reports the point estimate of τ_A from the number of applicants per labor market specification in (11), with the outcome N_{qd} in logs.

randomization inference p-value is 0.984, indicating that we cannot reject the sharp null of no impact of advertised P4P. The third column reports results for the TTC score regression with 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 (i.e. the status quo mapping from TTC scores to hiring probabilities). The fourth column reports results for the TTC score regression with weights $w_{iqd} = 1$ for the top $\hat{H}$ teachers in their application pool, and zero otherwise (i.e. a meritocratic hiring rule based on TTC scores alone). Here, we test for 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. Neither set of weights changes the conclusion from the second column: we cannot reject the sharp null of no impact of advertised P4P. The final column reports results for the (logged) application volume regression. Our estimate of τ_A is -0.040. The randomization inference p-value of 0.811, indicating that we cannot reject the sharp null of no impact of advertised P4P on application volumes.

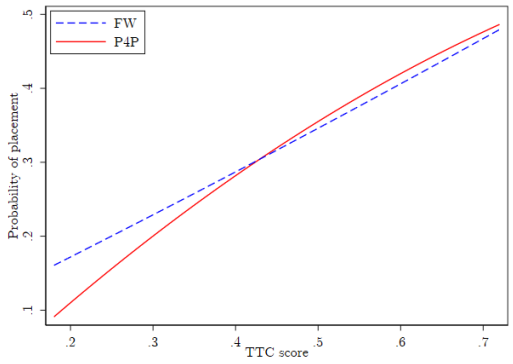
Supplementary analysis of placements

Interpretation of the effects of *advertised* P4P as purely a labor-supply response rests on the assumption that there are no changes to hiring patterns in response to treatment. This is testable in our setting, where interviews are not part of the hiring process, so that we have access to the full set of characteristics observed by

District Education Officers making hiring decisions.

A sufficient—but not necessary—test of the absence of a demand-side response would be to test whether the probability of hiring, as a function of observed applicant characteristic x_j, is the same under both P4P and FW advertisements.[46] We present a simple test of this in Figure C.1. There, we plot the empirical probability of hiring as a (quadratic) function of the *rank* of an applicant's TTC score within the set of applicants in their district.[47] It is clear from the figure that the predicted probabilities are similar across P4P and FW labor markets. A formal test confirms that the association between TTC score and the probability of hiring is not statistically significantly different across advertised treatment arms; i.e. there is no evidence of changes in hiring patterns in response to treatment.

Figure C.1: Probability of hiring as a function of TTC score, by advertised treatment arm



Notes: The figure illustrates estimated hiring probability as a (quadratic) function of the rank of an applicant's TTC final exam score within the set of applicants in their district.

[46] This condition is not necessary, because it is possible that the probability of offers being *accepted* by applicants is affected by the advertised contract associated with that post, even if applicants apply to jobs of both types and even if DEOs do not take contract offer types into account when selecting the individuals to whom they would like to make offers.

[47] We focus on within-applicant-pool ranks of TTC scores, rather than their unconditional values, because a rank-based offer rule seems the most plausible, and because this avoids confounds that might arise due to chance variation in applicant pool quality.

Appendix D Teacher value added

This section briefly summarises how we construct the measure of teacher value added for the placed recruits, referred to at the end of Section 4.1.

We follow the prior literature, most notably Kane and Staiger (2008) and Bau and Das (forthcoming). Denoting learning outcomes of student i in subject b, stream k of grade g, taught by teacher j in school s and round r, we express the data-generating process as:

$$y_{ibgjr} = \rho_{bgr} y_{i,r-1} + \mu_{bgr} + \lambda_s + \theta_j + \eta_{jr} + \varepsilon_{ibgjr}, \tag{12}$$

where θ_j is the time-invariant effect of teacher j: her value added. We allow for fixed effects by subject-grade-rounds, μ_{bgr}, and schools λ_s, estimating these within the model. We then form empirical Bayes estimates of TVA as follows.

1. Estimate the variance of the TVA, teacher-year, and student-level errors, $\theta_j, \eta_{jr}, \varepsilon_{ibgjr}$ respectively, from equation (12). Defining the sum of these errors as $v_{ibgjr} = \theta_j + \eta_{jr} + \varepsilon_{ibr}$: the last variance term can be directly estimated by the variance of student test scores around their teacher-year means: $\hat{\sigma}^2_\varepsilon = \text{Var}(v_{ibgjr} - \bar{v}_{jr})$; the variance of TVA can be estimated from the covariance in teacher mean outcomes across years: $\hat{\sigma}^2_\theta = \text{Cov}(\bar{v}_{jr}, \bar{v}_{j,r-1})$, where this covariance calculation is weighted by the number of students taught by each teacher; and the variance of teacher-year shocks can be estimated as the residual, $\hat{\sigma}^2_\eta = \text{Var}(v_{ibgjr}) - \hat{\sigma}^2_\theta - \hat{\sigma}^2_\varepsilon$.

2. Form a weighted average of teacher-year residuals $\bar{v}_{jr}$ for each teacher.

3. Construct the empirical Bayes estimate of each teacher's value added by multiplying this weighted average of classroom residuals, $\bar{v}_j$, by an estimate of its reliability:

$$\widehat{VA}_j = \bar{v}_j \left(\frac{\hat{\sigma}^2_\theta}{\text{Var}(\bar{v}_j)} \right) \tag{13}$$

   where $\text{Var}(\bar{v}_j) = \hat{\sigma}^2_\theta + (\sum_r h_{jr})^{-1}$, with $h_{jr} = \text{Var}(\bar{v}_{jr}|\theta_j)^{-1} = \left(\hat{\sigma}^2_\eta + \frac{\hat{\sigma}^2_\varepsilon}{n_{jr}} \right)^{-1}$.

Following this procedure, we obtain a distribution of (empirical Bayes estimates of) teacher value added for placed recruits who applied under advertised FW. The Round 2 point estimate from the student learning model in Equation (3) would raise a teacher from the 50th to above the 76th percentile in this distribution. Figure

4 plots the distributions of (empirical Bayes estimates of) $\theta_j + \eta_{jr}$ separately for $r = 1, 2$, and for recruits appling under advertised FW and advertised P4P.

It is of interest to know whether the measures of teacher ability and intrinsic motivation that we use in Section 4.1 are predictive of TVA. This is undertaken in Table D.1, where TVA is the estimate obtained pooling across rounds and treatments.[48] Interestingly, the measure of teacher ability that we observe among recruits at baseline, Grading Task IRT score, *is* positively correlated with TVA (rank correlation of 0.132, with a p-value of 0.039). It is also correlated with TTC final exam score (rank correlation of 0.150, with a p-value of 0.029). However, neither the measure that districts have access to at the time of hiring, TTC final exam score, nor the measure of intrinsic motivation that we observe among recruits at baseline, DG share sent, is correlated with TVA.

Table D.1: Rank correlation between TVA estimates, TTC scores, Grading Task IRT scores, and Dictator Game behavior among new recruits

	TVA	TTC score	Grading task
TTC score	-0.087	.	.
	(0.178)		
Grading task	0.132	0.150	.
	(0.039)	(0.029)	
DG share sent	-0.078	0.062	-0.047
	(0.203)	(0.349)	(0.468)

Notes: The table provides rank correlations and associated p-values (in parentheses) for relationships between recruits' teacher value added and various measures of skill and motivation: TTC final exam scores, baseline Grading Task IRT scores, and baseline Dictator Game share sent. We obtain the empirical Bayes estimate of TVA from θ_j estimated in the school fixed-effects model in equation (12).

[48] We obtain qualitatively similar results for the FW sub-sample, where TVA cannot be impacted by treatment with P4P.

Online Appendix References

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

Leaver, Clare, Renata Lemos, and Daniela Scur, "Measuring and explaining management in schools: New approaches using public data," CEPR Discussion Paper DP14069 October 2019.

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

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