American Economic Review 2021, 111(7): 2213–2246
https://doi.org/10.1257/aer.20191972

Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools[†]

By Clare Leaver, Owen Ozier, Pieter Serneels, and Andrew Zeitlin*

This paper reports on 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 were 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. By the second year of the study, the within-year effort effect of P4P was 0.16 standard deviations of pupil learning, with the total effect rising to 0.20 standard deviations after allowing for selection. (*JEL* C93, I21, J23, J33, J41, J45, O15)

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, Friedman, and Rockoff 2014a, b) as well as social returns through improved labor market skills that drive economic growth

*Leaver: Blavatnik School of Government, University of Oxford and CEPR (email: clare.leaver@bsg.ox.ac.uk); Ozier: Department of Economics, Williams College, World Bank Development Research Group, BREAD, and IZA (email: owen.ozier@williams.edu); Serneels: School of International Development, University of East Anglia, EGAP, and IZA (email: p.serneels@uea.ac.uk); Zeitlin: McCourt School of Public Policy, Georgetown University, and CGD (email: andrew.zeitlin@georgetown.edu). Esther Duflo was the coeditor for this article. 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 the three anonymous referees, 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 UK 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. Leaver is grateful for the hospitality of the Toulouse School of Economics, 2018–2019. Research was conducted under Rwanda Ministry of Education permit number MINEDUC/S&T/308/2015 and received IRB approval from the Rwanda National Ethics Committee (protocol 00001497) and from Innovations for Poverty Action (protocol 1502). This study is registered as AEA RCT Registry ID AEARCTR-0002565 (Leaver et al. 2018). 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.

[†]Go to https://doi.org/10.1257/aer.20191972 to visit the article page for additional materials and author disclosure statements.

(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* (hereafter P4P). These compensation schemes typically reward teacher inputs such as presence and conduct in the classroom, teacher value added (TVA) based on student learning, or both (see, for example, 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 P4P divides opinion. Critics, drawing upon public administration, social psychology, and behavioral economics, argue that P4P could dampen the effort of workers (Deci and Ryan 1985, Krepps 1997, Bénabou and Tirole 2003). 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 unmotivated 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 P4P 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 P4P. 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 P4P on the recruitment, effort, and retention of civil servant teachers.

At the center of this study is a P4P contract, designed jointly with the Rwanda Education Board (REB) 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 (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 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 reassigned 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 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).

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). Averaging over the two years of the study, the within-year effort effect of P4P was 0.11 standard deviations of pupil learning, and for the second year alone, the within-year effort effect of P4P was 0.16 standard deviations. 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 sizable 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 P4P.

To sum up, by the second year of the study, we estimate the within-year effort effect of P4P to be 0.16 standard deviations of pupil learning, with the total effect rising to 0.20 standard deviations after allowing for selection. Despite evidence of lower intrinsic motivation among those recruited under P4P, these teachers were at least as effective in promoting learning as those recruited under FW. These results support the view that P4P can improve effort while also allaying fears of harmful

effects on selection. Of course, we have studied a two-year intervention—impacts of a long-term policy might be different, particularly if P4P influences individuals' early-career decisions to train as a teacher.

Our findings bring new experimental results on P4P 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, Finan, and Rossi (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 prosocial applicants for village promoter roles, resulting in lower effort and retention. And in Zambia, Ashraf et al. (2020) find that emphasis on career-track motivations for community health work, while attracting some applicants who were less prosocial, resulted in hires of equal prosociality and greater talent overall, leading to improvements in a range of health outcomes. By studying P4P and 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 P4P is of interest in high-income contexts as well. In the United States, 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 (Neal 2011; Chetty, Friedman, and Rockoff 2014b; Rothstein 2015) or examined the role of teachers' outside options in labor supply (Chingos and West 2012). Recent work has examined Washington, DC'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 less likely to quit than their low value-added counterparts. Our prospective, experimental study of P4P contributes to this literature methodologically but also substantively since the Rwandan labor market shares important features with high-income contexts.[1]

While our paper is not the first on the broader topic of incentive-based contracts for teachers,[2] we go to some length to address two challenges thought to be important for policy implementation at scale. One is that the structure of the incentive should 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

[1] 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, Olken, and Pande 2017). The 2017 Rwanda Labour 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 FRw43,431, while 15 percent were in nonteaching jobs earning a higher average monthly salary of FRw56,347—a *private sector* premium of close to 30 percent (National Institute of Statistics of Rwanda 2017).

[2] See, for example, Imberman (2015) and Jackson, Rockoff, and Staiger (2014), who provide a review.

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, Romero, and Schipper (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 (Leaver et al. 2018), 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, for example, Olken (2015).[3] 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).[4] For an experimental study in which one important dimension of variation occurs at the labor-market level, and so is potentially limited in power, the gains from these specification choices are particularly important. The results reported in our pre-analysis plan demonstrate that with specifications appropriately chosen, the study design is well powered, such that even null effects would be of both policy and academic interest.

In the remainder of the paper, Sections I and II describe the study design and data, Sections III and IV report and discuss the results, and Section V concludes.

I. Study Design

A. *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.[5] 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. 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

[3] We have not found prior examples of such blinding in economics. Humphreys, Sanchez de la Sierra, and van der Windt (2013) argue for, and undertake, a related approach with partial endline data in a political science application.

[4] 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.

[5] "Upper primary" refers to grades four, five, and six; schools typically include grades one through six.

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.

Given this institutional setting, we can think of district-by-subject-family pairs as *labor markets*. The subject-family boundaries of these labor markets are rigid; within each district, TTC degree holders are considered for jobs in pools alongside others with the same qualification. The district boundaries may be more porous, though three quarters of the new teaching jobs in our study were filled by recruits living within the district at the time of application. Since this is the majority of jobs, we proceed by treating these labor markets as distinct for our primary analysis and provide robustness checks for cross-district applications in online Appendix C.[6]

There are 18 such labor markets in our study.[7] 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 nonteaching, salaried employment (National Institute of Statistics of Rwanda 2017). This is not because the teacher labor market is tight: nationwide, close to a quarter of vacancies created by a teacher leaving a school remain unfilled in the following school year (Zeitlin 2021). A more plausible explanation is that the recent graduates in the outside sector earned a premium of close to 30 percent, making occupational choice after TTC a meaningful decision.

B. *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 P4P contract, which paid FRw100,000 (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.[8] 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).[9]

[6] As we note in the online Appendix, cross-district applications would not lead us to find a selection effect where none existed, but we might overstate the magnitude of any selection effect.

[7] 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.

[8] The exchange rate on January 1, 2016, was FRw734 to $1, so the FRw100,000 bonus was worth roughly $136.

[9] 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. The data used to construct this measure, and the measures of teachers' inputs, are described in Sections IIC and IID, respectively, and we explain the adaptation of the Barlevy-Neal measure of learning outcomes to a repeated cross-section of pupils in online Appendix D.

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 FW contract that paid FRw20,000 to all upper-primary teachers. This bonus was paid at the same time as the performance award in the P4P contract.

Although P4P contracts based on a composite metric of teacher inputs and student performance have been used in a number of policy settings in the United States (Imberman 2015, Stecher et al. 2018), such contracts have been relatively less studied in low- and middle-income countries. In their comprehensive review, Glewwe and Muralidharan (2016) discuss several evaluations of teacher incentives based on student test scores or attendance checks but none based on a combination of both. After extensive discussions with REB about what would be suitable in this policy setting, a decision was made to use the P4P contract described previously, based on a composite metric.

Design Overview.—The design, summarized visually in Figure A.1 in online Appendix A, draws on a two-tiered experiment, as used elsewhere (see Karlan and Zinman 2009; Ashraf, Berry, and Shapiro 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 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 P4P contracts on teacher performance, whereas comparisons along the diagonal of *a*–*d* are informative about the total effect of such contracts, along both margins.

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 seven of these labor markets can be thought of as being in a "P4P only" advertised treatment, seven in a "FW only" advertised treatment, and four in a "mixed" advertised treatment.[10]

[10] 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.

		Advertised	
		FW	P4P
Experienced	FW	*a*	*b*
	P4P	*c*	*d*

Figure 1. Treatment Groups among Recruits Placed in Study Schools

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.[11] 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. 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.[12] Following a full baseline survey, 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 the 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 were eventually offered another contract might have experienced disappointment (or other negative feelings) that then had a causal impact on their behavior. To mitigate this concern, all new recruits were offered an *end-of-year retention bonus* of FRw80,000 on top of their school-randomized P4P or FW contract. An individual who applied under advertised P4P in the hope of receiving FRw100,000 from the scheme but was subsequently re-randomized to experienced FW was therefore still eligible to

[11] Details of the promotional materials used in this campaign are provided in online Appendix E.

[12] Because schools could receive multiple recruits, for different teaching specializations, it was possible for enrolled schools to contain two recruits who had experienced distinct advertised treatments. Recruits hired under the mixed advertisement treatment, and the schools in which they were placed, also met our enrollment criteria. These were similarly re-randomized to either experienced P4P or experienced FW in the second-tier randomization.

receive FRw100,000 (FRw20,000 from the FW contract plus FRw80,000 as a retention bonus). Conversely, an individual who applied under advertised FW safe in the knowledge of receiving FRw20,000 from the scheme but who was subsequently re-randomized to experienced P4P was still eligible for at least FRw80,000. 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. To ensure that teachers in P4P schools understood the new contract, we held a compulsory half-day briefing session in every P4P school to explain the intervention. This session was conducted by a team of qualified enumerators and District Education Office staff, who themselves received three days of training from the principal investigators in cooperation with IPA. Online Appendix E reproduces an extract of the English version of the enumerator manual, which was piloted before use. The sessions provided ample space for discussion and made use of practical examples. Teachers' understanding was tested informally at the end of the session. We also held a comparable (but simpler) half-day briefing session in every FW school.

C. *Hypotheses*

Precommitment 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 and online Appendix B, 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 $w - \left(e^2 - \tau e\right)$, 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.[13] Effort generates a performance metric $m = e\theta + \varepsilon$, 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

[13] See Delfgaauw and Dur (2008) for a related approach to modeling differential worker motivation across sectors.

in Figure 1. The timing is as follows. Teacher vacancies are advertised as either P4P or FW. The individual, of type (τ, θ), applies to either a teaching job or an outside job. Employers hire, at random, from the set of (τ, θ) types that apply. Thereafter, contracts are re-randomized. 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:

(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 that 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: that is, to hypothesis II via θ, hypothesis III via τ, and hypotheses IV to VI via the effect of θ and τ on performance.[14] 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 hypotheses 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, Sanchez de la Sierra, and van der Windt (2013) and Olken (2015). Researchers can use actual outcome data with the treatment variable scrambled or removed to estimate

[14] When mapping the theory to our empirical context, we distinguish between these hypotheses for two reasons: we have better data for placed recruits because we were able to administer detailed survey instruments to this well-defined subsample, and for placed recruits we can identify the advertised treatment effect from student learning outcomes, avoiding the use of proxies for (τ, θ). A further consideration is that the impact of advertised treatment might differ between placed recruits and applicants due to labor-demand effects. We discuss this important issue in Section IV.

specifications in "mock" data. This permits navigation of an otherwise intractable "analysis tree." They can also improve 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.[15]

The second set of insights pertains 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.[16]

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 yield lower statistical power than would a Kolmogorov-Smirnov (KS) test of the equality of two distributions. Over a range of simulations, the KS test had between one and four times the power of OLS. We therefore committed to KS (over OLS and two other alternatives) as our primary test of this hypothesis.[17] This prediction is borne out in Table C.1 in online Appendix C.[18]

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 school level, we had many 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 could produce a 20 to 25 percent narrower confidence interval than in a simple benchmark specification.[19] Comparing Table 3 to Table A.4 in online Appendix A, this was substantively borne out.[20]

[15] 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 that 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 (Leaver et al. 2018, AEARCTR-0002565) is accompanied by IPA's letter specifying the date after which treatment was unblinded.

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

[17] This refers to Leaver et al. (2018), table C.1, comparing the first and third rows.

[18] 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.

[19] This refers to Leaver et al. (2018), table C.3, comparing row 12 to row 1.

[20] For the pooled advertised treatment effect, the precommitted random effects model yields a confidence interval that is 67 percent as wide as the interval from OLS: a gain in precision commensurate with increasing the sample

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 Table A.1 in online Appendix A. We also included a small number of secondary tests based on different outcomes, samples, and/or specifications. In Section III, we report results for every primary test; secondary tests are in Section III or in an online Appendix. To aid interpretation, we also include a small amount of supplementary analysis that was not discussed in the pre-analysis plan—for example, impacts of advertised P4P on teacher attributes beyond observable skill and intrinsic motivation and estimates from a TVA model—but are cautious and make clear when this is post hoc.

II. 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 candidates' application decisions, to the set (and attributes) of candidates hired into schools, to the learning outcomes that they deliver, and, finally, to the teachers' 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.[21] Our understanding of these data informs our choices of specification for analysis, as discussed in detail in the pre-analysis plan.

A. *Applications*

Table 1 summarizes the applications for the newly advertised jobs, submitted in January 2016, across the six districts.[22] Of the 2,184 applications, 1,962 come from candidates with a TTC degree. We term these *qualified* since 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 2,184 applications come from 1,424 unique individuals, of whom 1,246 have a TTC qualification. The majority (62 percent) of qualified applicants complete only one application, with 22 percent applying to two districts and 16 percent applying to three or more. Multiple applications are possible but not the norm, most likely because each district requires its own exam. Of those applying twice, 92 percent applied to adjacent pairs of districts. In Online Appendix C, we use this geographical feature of applications to test for cross-district labor-supply effects and fail to reject the null that these effects are zero.

size by 125 percent. The gain in precision for the pooled experienced treatment effect is smaller and commensurate with increasing the sample size by 22 percent.

[21] All data generated by the study and used in this paper are made available in the replication materials (Leaver et al. 2020).

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

Table 1—Application Characteristics, by District

	Gatsibo	Kayonza	Kirehe	Ngoma	Nyagatare	Rwamagana	All
Applicants	390	310	462	380	327	315	2,184
Qualified	333	258	458	364	272	277	1,962
Has TTC score	317	233	405	337	260	163	1,715
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	27.25	26.98	27.50	27.33

B. *Teacher Attributes*

During February and March 2016, we visited schools soon after they were enrolled in the study to collect baseline 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 their eligibility 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, 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 toward students, taking inspiration from the work of Ashraf, Bandiera, and Jack (2014), we used a framed version of the *dictator game* (Eckel and Grossman 1996).[23] Teachers were given FRw2,000 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 FRw200. Teachers could decide to allocate any amount, from FRw0 to all FRw2,000, which would supply ten randomly chosen students with a packet.

We also asked teachers to undertake a *grading task* that measured their mastery of the curriculum in the main subject that they teach.[24] Teachers were asked to grade a student examination script and had five 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.

[23] Previous work shows the reliability of the dictator game as a measure of other-regarding preferences related to intrinsic motivation (Brock, Lange, and Leonard 2016; Banuri and Keefer 2016; Deserranno 2019).

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

C. *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 four, five, and six. 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. 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. The tests were piloted extensively; they have no ceiling effects, while floor effects are limited.[25] In each round, we randomly sampled a subset of students from each grade to take the test. In year 1, both 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 ensured that the sampling protocol did not create incentives for strategic exclusion of students. In year 2, students were assessed only at the end of the year 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 eight streams, we sampled five pupils from all streams. In round 1, we sampled ten pupils from each stream: five pupils retained from the baseline (if the stream was sampled at baseline) and five 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 ten pupils from each stream using the listing for that year.[26]

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, which further ensured consistency in test administration. Individual student test results were kept confidential from teachers, parents, head teachers, and

[25] Test scores are approximately normally distributed with a mean of close to 50 percent of questions answered correctly. A validation exercise of the test at baseline found its scores to be predictive of the national Primary Leaving Exam scores (both measured in school averages).

[26] Consequently, the number of pupils assessed in year 2 who were also 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.

Ministry of Education officials and have been used only 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.

D. *Teacher Inputs*

We collected data on several dimensions of teachers' inputs into the classroom. This was undertaken in only P4P schools during year 1 and in both P4P and FW schools in year 2. This composite metric is based on three teacher input measures (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.[27] FW schools also received an unannounced visit in year 2 at the same time as the P4P schools. Table A.2 in online Appendix A 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.

Lesson *preparation* is defined as the planning involved in daily lessons and is measured through a review of teachers' weekly lesson plans. Prior to any spot checks, teachers in grades four, five, and six in P4P schools were reminded how to fill out a lesson plan in accordance with REB guidelines. Specifically, SEOs provided teachers with a template to record their lesson preparation, focusing on three key components of a lesson: the lesson objective, the instructional activities, and the types of assessment to be used. A "hands-on" session enabled teachers to practice using this template. 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 to impact student learning. We collaborated with both the Ministry of Education and REB to develop a monitoring instrument to measure teacher pedagogy through

[27] Training of SEOs took place over two days. Day one 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 using the enumerator manual in online Appendix E. During day two, 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.

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 United States (Danielson 2007). The observer evaluated the teachers' effective use of 21 different activities over the course of a full 45-minute lesson. 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. 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.

E. *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.[28]

III. Results

Our two-tiered experiment allows us to estimate impacts of P4P 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 IIIA and IIIB, respectively. Of course, the long-run effects of P4P will depend on not only selection in but also selection *out*, as well as the dynamics of the behavioral response on the part of teachers who stay. We address dynamic issues in Section IIIC and postpone a substantive discussion of results until Section IV. All statistical tests are conducted via randomization inference with 2,000 permutations of the relevant treatment.

A. *Compositional Margin of P4P*

We study three types of compositional effects of P4P: 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 IC, 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 TTC final exam score since this is the only consistently measured quality-related characteristic we observe for all applicants.

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

Table 2—Baseline Characteristics and Balance of Experienced P4P Assignment

	Control mean [SD]	Experienced P4P (p-value)	Observations
Panel A. School attributes			
Number of streams	9.99	−0.10	164
	[4.48]	(0.881)	
Number of teachers	20.47	0.56	164
	[8.49]	(0.732)	
Number of new recruits	1.94	0.13	164
	[1.30]	(0.505)	
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 recruit attributes			
Female	0.36	−0.02	242
	[0.48]	(0.770)	
Age	25.82	−0.25	242
	[4.05]	(0.616)	
Dictator game share sent	0.28	−0.04	242
	[0.33]	(0.450)	
Grading task score	−0.24	0.12	242
	[0.93]	(0.293)	
Panel C. Pupil learning assessments			
English	−0.00	0.04	13,826
	[1.00]	(0.551)	
Kinyarwanda	−0.00	0.05	13,831
	[1.00]	(0.292)	
Mathematics	0.00	−0.00	13,826
	[1.00]	(0.950)	
Science	−0.00	0.03	13,829
	[1.00]	(0.607)	
Social Studies	−0.00	0.02	13,829
	[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 (with standard deviations in brackets); the second column presents estimated differences between FW and P4P schools (with randomization inference p-values in parentheses). The sample in panel B consists of new recruits placed in upper-primary classrooms at baseline who undertook the lab-in-the-field exercises. 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.

Our primary test uses a 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

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

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 p-value is then the share of such test statistics larger in absolute value than the statistic calculated 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 based on inversion of the randomization inference test rules out additive treatment effects outside of the range $[-0.020, 0.020]$. (The OLS estimate of this additive treatment effect is -0.001, as reported in Table C.1 in online Appendix C.) We therefore conclude that there was no meaningful impact of advertised P4P on the TTC final exam scores of applicants.[29]

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 scores 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.—Along the lines suggested by Dal Bó and Finan (2016), we explore whether institutions can attract the most capable or the most intrinsically motivated into public service. 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 IIB. 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 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 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.[30] 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, $T^A \in \mathcal{T}^A$.

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

[30] 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.

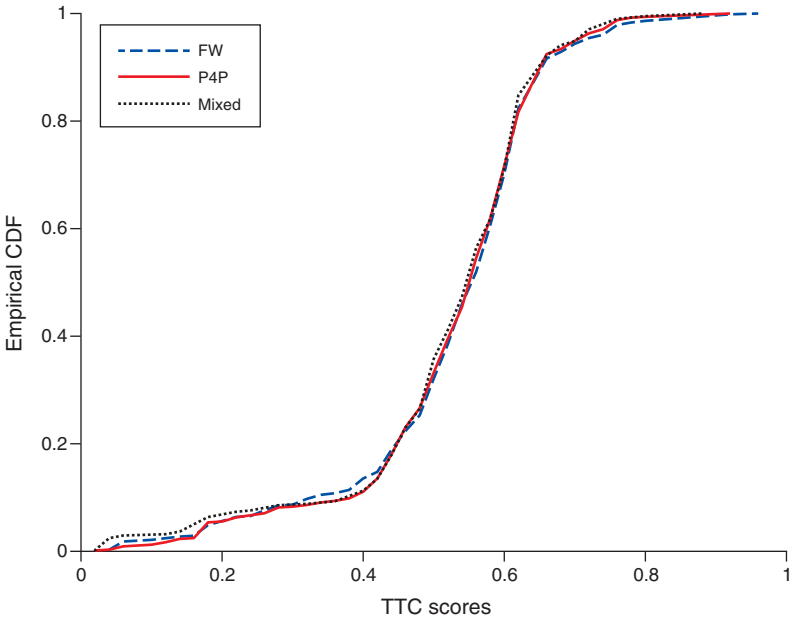

FIGURE 2. DISTRIBUTION OF APPLICANT TTC SCORE, BY ADVERTISED TREATMENT ARM

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

Before reporting these t-statistics, it is instructive to view the data graphically. Figure 3, panel A, shows the distribution of grading task IRT scores and Figure 3, panel B, 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 -0.184, with a (randomization inference) p-value of 0.367. In the dictator game share sent specification, our estimate of τ_A is -0.100, with a p-value of 0.029. 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.

We chose not to include additional teacher attributes in the theoretical model and in the list of pre-specified hypotheses to avoid multiple hypothesis testing concerns. Notwithstanding this decision, we did collect additional data on placed recruits at baseline, meaning that we can use our two-tiered experimental design to conduct further *exploratory* analysis of the impact of advertised P4P. Specifically, we estimate regressions of the form given in equation (2) for four additional teacher attributes: age, gender, risk aversion, and an index capturing the big five personality traits.[31] Results are reported in Table A.3 in online Appendix A, with details of the

[31] Here we follow Dal Bó, Finan, and Rossi (2013), who measure the risk preferences and big five personality traits of applicants for civil service jobs in Mexico, and Callen et al. (2020), who study the relevance of big five

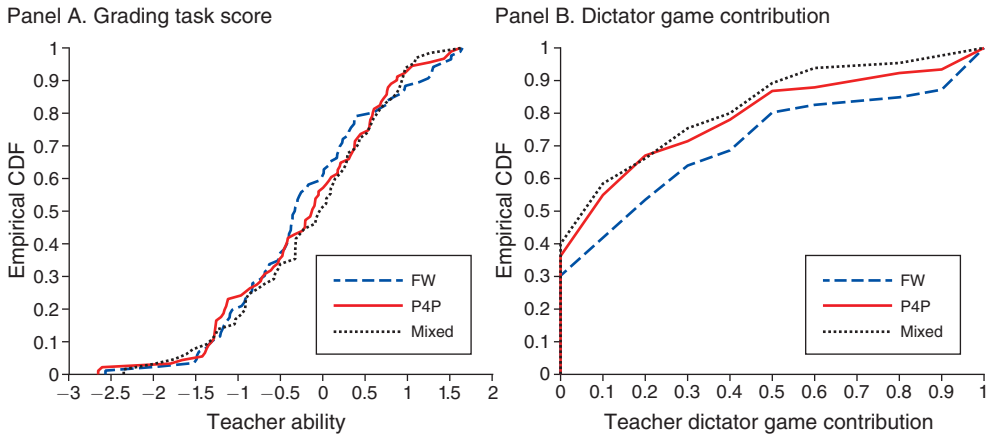

FIGURE 3. DISTRIBUTION OF PLACED RECRUIT ATTRIBUTES ON ARRIVAL, BY ADVERTISED TREATMENT ARM

Notes: In panel A, the t-statistic for a difference in mean grading task IRT score across the P4P and FW treatments is -0.184, with a p-value of 0.367. In panel B, the t-statistic for a difference in mean dictator game share sent across the P4P and FW treatments is -0.100, with a p-value of 0.029.

variable construction provided in the table note. We are unable to reject the sharp null of no advertised P4P treatment effect for any of these exploratory outcomes.

Student Learning Induced by Placed Recruits.—The skill and motivation of placed recruits on arrival are policy relevant insofar as these attributes translate into teacher effectiveness. To assess this, we combine experimental variation in the advertised contracts to which recruits applied, with the second-stage randomization in the 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. 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 denoted by T^E_s. We pool data across the two years of intervention to estimate a specification of the type

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

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

personality traits for the performance of health workers in Pakistan.

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). Drawing on the pseudo panel of student outcomes, 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 average of the within-year effect of advertised P4P on pupil learning in year 1 and the within-year effect of advertised P4P on pupil learning in year 2.[32]

The theoretical model of online Appendix B, as well as empirical evidence from other contractual settings (Einav et al. 2013), suggests that P4P 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

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

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*). If τ_{AE} is not zero, then this interacted model yields the more policy relevant estimands (Muralidharan, Romero, and Wüthrich 2019). Noting the distinction between estimands and test statistics (Imbens and Rubin 2015), we prespecified the pooled coefficient τ_A from equation (3) as the primary test statistic for the presence of compositional effects. Our simulations, using blinded data, show that this pooled test is better powered under circumstances where the interaction term, τ_{AE}, is small.

We estimate equations (3) and (4) by a linear mixed effects model, allowing for normally distributed random effects at the student-round level.[33] Randomization inference is used throughout. To do so, we focus on the distribution of the estimated *z*-statistic (that is, 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 nonsharp null that the coefficient is equal to zero (DiCiccio and Romano 2017). Inference for τ_A is undertaken by permutation of the advertised treatment, $T^A \in \mathcal{T}^A$, while inference for τ_E likewise proceeds by permuting the experienced treatment $T^E \in \mathcal{T}^E$. To conduct inference about the interaction term, τ_{AE} in equation (4), we simultaneously permute both dimensions of the treatment, considering pairs $\left(T^A, T^E\right)$ from the set $\mathcal{T}^A \times \mathcal{T}^E$.

[32] We focus on *within-year* impacts because there is not a well-defined cumulative treatment effect. Individual students receive differing degrees of exposure to the advertised treatments depending on their paths through streams (and hence teachers) over years 1 and 2.

[33] 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 maximize 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 Table A.4 in online Appendix A. 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 (τ_A)	0.01	−0.03	0.04
	[−0.04, 0.08]	[−0.06, 0.03]	[−0.05, 0.16]
	(0.75)	(0.20)	(0.31)
Experienced P4P (τ_E)	0.11	0.06	0.16
	[0.02, 0.21]	[−0.03, 0.15]	[0.04, 0.28]
	(0.02)	(0.17)	(0.00)
Experienced P4P × incumbent (λ_E)	−0.06	−0.05	−0.09
	[−0.20, 0.07]	[−0.19, 0.11]	[−0.24, 0.06]
	(0.36)	(0.54)	(0.27)
Model B. Interactions between advertised and experienced contracts			
Advertised P4P (τ_A)	0.01	−0.02	0.03
	[−0.05, 0.14]	[−0.06, 0.07]	[−0.05, 0.21]
	(0.46)	(0.62)	(0.22)
Experienced P4P (τ_E)	0.12	0.06	0.18
	[0.05, 0.25]	[−0.01, 0.19]	[0.08, 0.33]
	(0.01)	(0.10)	(0.00)
Advertised P4P × experienced P4P (τ_{AE})	−0.03	−0.01	−0.04
	[−0.17, 0.09]	[−0.15, 0.10]	[−0.22, 0.13]
	(0.51)	(0.65)	(0.58)
Experienced P4P × incumbent (λ_E)	−0.08	−0.05	−0.11
	[−0.31, 0.15]	[−0.30, 0.18]	[−0.36, 0.14]
	(0.43)	(0.56)	(0.38)
Observations	154,594	70,821	83,773

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 IIC.

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 the difference is not statistically 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 the second year of the study, the within-year compositional effect of P4P was 0.04 standard deviations of pupil learning. OLS estimates of this effect are larger, at 0.08 standard deviations, with a p-value of 0.10, as shown in online Appendix Table A.4.

For the purposes of interpretation, it is useful to recast the data in terms of TVA. As detailed in online Appendix D, we do so by estimating a 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 fiftieth to above the seventy-third percentile

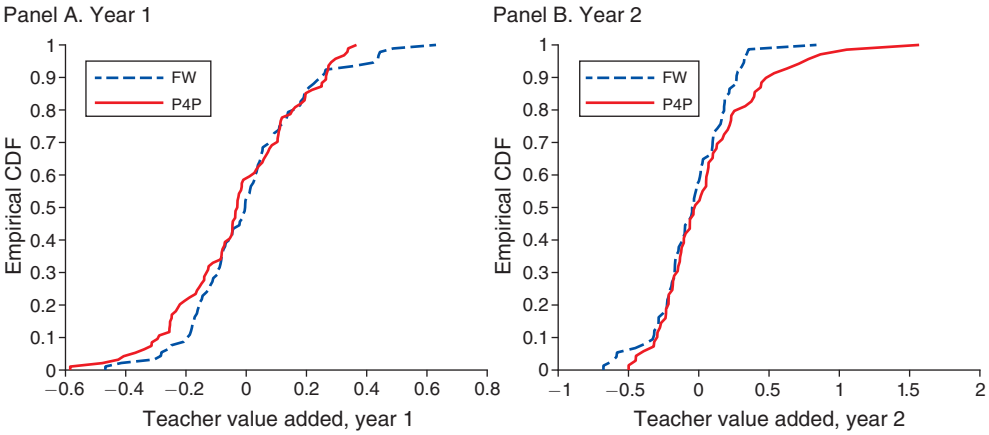

FIGURE 4. TVA AMONG RECRUITS, BY ADVERTISED TREATMENT AND YEAR

Notes: The figures plot distributions of TVA 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 TVA (not prespecified).

in the distribution of (empirical Bayes estimates of) TVA for placed recruits who applied under FW. The TVA model also reveals the impact of advertised P4P on the distribution of teacher effectiveness. Figure 4, panel B, shows that the distribution of TVA 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.

B. *Effort Margin of P4P*

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 $\tau_E + \tau_{AE}$.[34]

Results are presented in Table 3. Pooling across years, the within-year effect of experienced P4P is 0.11 standard deviations of pupil learning (model A, second row). The randomization inference p-value is 0.02, 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 the second year of the study, the within-year effort effect of P4P was 0.16 standard deviations of pupil learning.[35]

To put this in perspective, we compare the magnitude of this effort effect to impacts in similar studies in the United States and beyond. Sojourner, Mykerezi, and West (2014) study P4P schemes in Minnesota, typically based on a composite metric of subjective teacher evaluation and student performance, and find an effect of 0.03 standard deviations of pupil learning. Dee and Wyckoff (2015) study a high-stakes incentive over a composite metric in Washington, DC, and find effects consistent with those of Sojourner, Mykerezi, and West (2014) in terms of the implied magnitude of effects on pupil learning. Glewwe and Muralidharan (2016) review a range of studies including several in Benin, China, India, and Kenya that employ incentives for either students or teachers based solely on student performance; effect sizes are larger, typically above 0.2 standard deviations of pupil learning. Our effort effect falls within this range and is of a comparable magnitude to the impact in Duflo, Hanna, and Ryan (2012), who study incentives for teacher attendance in India.

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. 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 the teacher level (hypothesis VI). For these tests, we use the following specifications:

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

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

for the metric of teacher j with qualification q in school s of district d, as observed in posttreatment 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).[36] A linear

[34] We are grateful to a referee for highlighting a further interpretation: τ_E in model B is the policy-relevant estimate of experienced P4P at the start of any unexpected transition to P4P, while $\tau_E + \tau_{AE}$ is the policy-relevant estimate for that effect slightly further into a transition—the effect of P4P on a cohort anticipating P4P.

[35] Across all specifications, the interaction term between experienced P4P and an indicator for incumbent teachers is negative, though statistically insignificant, and smaller in magnitude than the direct effect of experienced P4P, implying a weaker—though still positive—effect of P4P on incumbents in point-estimate terms.

[36] 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

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, based on pooled data.[37] 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 subcomponent. 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 sizable in economic terms given that baseline teacher presence was already nearly 90 percent. Recruits who experienced P4P were 0.10 points more effective in their classroom practices than recruits receiving FW, although it is possible that this improvement occurred only during the observation. We find no evidence of impacts on lesson planning.

C. *Dynamic Effects*

Our two-tiered experiment was designed to evaluate the impact of P4P 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 be driven only 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 2011b). 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 hypothesized 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.[38]

are therefore not included as independent variables.

[37] As discussed in Section IID, FW schools received unannounced visits to measure teacher inputs only in year 2.

[38] We emphasize 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.

Table 4—Estimated Effects on Dimensions of the Composite "4P" Performance Metric

	Summary metric (1)	Preparation (2)	Presence (3)	Pedagogy (4)	Pupil learning (5)
Model A. Direct effects only					
Advertised P4P (τ_A)	−0.04	0.07	0.00	0.03	−0.02
	[−0.09, 0.01]	[−0.13, 0.32]	[−0.05, 0.07]	[−0.06, 0.10]	[−0.08, 0.02]
	(0.11)	(0.40)	(0.93)	(0.42)	(0.27)
Experienced P4P (τ_E)	0.23	0.02	0.08	0.10	0.09
	[0.19, 0.28]	[−0.13, 0.16]	[0.02, 0.14]	[−0.00, 0.21]	[0.03, 0.15]
	(0.00)	(0.84)	(0.01)	(0.05)	(0.00)
Experienced P4P	0.03	0.07	−0.01	0.07	−0.00
× incumbent (λ_E)	[−0.01, 0.07]	[−0.03, 0.18]	[−0.06, 0.05]	[−0.01, 0.16]	[−0.04, 0.03]
	(0.10)	(0.17)	(0.70)	(0.11)	(0.86)
Model B. Interactions between advertised and experienced contracts					
Advertised P4P (τ_A)	−0.03	0.16	−0.01	0.12	−0.01
	[−0.12, 0.05]	[−0.11, 0.48]	[−0.16, 0.17]	[−0.27, 0.55]	[−0.12, 0.11]
	(0.42)	(0.19)	(0.86)	(0.44)	(0.91)
Experienced P4P (τ_E)	0.22	−0.00	0.08	0.17	0.08
	[0.15, 0.29]	[−0.26, 0.25]	[−0.01, 0.16]	[−0.05, 0.38]	[0.00, 0.16]
	(0.00)	(0.97)	(0.07)	(0.12)	(0.04)
Advertised P4P	−0.02	−0.11	0.02	−0.11	−0.03
× experienced P4P (τ_{AE})	[−0.11, 0.07]	[−0.45, 0.23]	[−0.12, 0.16]	[−0.45, 0.24]	[−0.15, 0.08]
	(0.65)	(0.53)	(0.69)	(0.53)	(0.64)
Experienced P4P	0.05	0.09	−0.01	0.00	0.00
× incumbent (λ_E)	[−0.01, 0.10]	[−0.07, 0.26]	[−0.09, 0.07]	[−0.13, 0.14]	[−0.05, 0.06]
	(0.07)	(0.27)	(0.82)	(0.96)	(0.90)
Observations	3,996	2,514	3,455	2,136	3,049
FW recruit mean	0.49	0.65	0.89	1.98	0.48
SD	(0.22)	(0.49)	(0.31)	(0.57)	(0.27)
FW incumbent mean	0.37	0.50	0.87	2.05	0.45
SD	(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 deviation) 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 only the second year. Outcomes are constructed at the 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.

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 year 2—that is, after experiencing P4P in year 1 but before the performance awards were announced. To do so, we use a linear probability model of the form

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

where $employed_{iqd2}$ is an indicator for whether 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.

As column 1, Table 5 reports, our estimate of τ_E is zero with a randomization inference p-value of 0.94. There is no statistically significant impact of experienced

Table 5—Retention of Placed Recruits

	(1)	(2)	(3)
Experienced P4P	0.00	−0.04	−0.08
	(0.94)	(0.42)	(0.24)
Interaction		−0.05	0.15
		(0.39)	(0.37)
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 parentheses. Randomization inference is conducted on the associated *t*-statistic. In each column, the outcome is an indicator for whether the teacher is still employed at the start of year 2. The mean of this dependent variable for FW recruits is 0.80. In the second column, the specification includes an interaction of experienced treatment with the teacher's baseline grading task IRT score (not de-meaned); in the third column, the interaction is with the teacher's share sent in the baseline framed dictator game (again, not de-meaned). All specifications include controls for districts and subjects of teacher qualification.

P4P on retention of recruits; the retention rate is practically identical (at around 80 percent) among recruits experiencing P4P and those experiencing FW.

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 five-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 online Appendix A Table A.5 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 0.39.

Of course, a retention rate of 80 percent implies 20 percent attrition from year 1 to year 2, which is nonnegligible. 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 teachers' performance on the baseline grading task in the primary subject they teach to obtain an IRT estimate of their ability in this subject, denoted z_i, and estimate an interacted model of the form

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

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 0.39. 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 reestimate the interacted model in equation (8), replacing z_i with x_i. As column 3, Table 5 reports, our

estimate of ζ in this specification is 0.15, with a randomization inference p-value of 0.37. 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 0.68, with a randomization inference p-value of 0.57. 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 the *rising* effects of experienced P4P from year 1 to year 2 are driven by *positive* changes in our measures of within-retained-recruit teacher skill or intrinsic motivation.[39]

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 at endline: job satisfaction, likelihood of leaving, and positive/negative affect.[40] As online Appendix Table A.5 shows, there is no statistically significant impact of experienced P4P on any of these measures.

Further substantiating this point, online Appendix A Table A.6 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?"[41] The proportion giving a favorable answer exceeds 75 percent in every study arm. In terms of Figure 1, group *a* (recruits who both applied for and experienced FW) had the most negative view of P4P, while group *c* (who applied for FW but experienced P4P) had the most positive view. Hence it seems that it was the idea, rather than the reality, of P4P that was unpopular with (a minority of) recruits.[42]

[39] Although repeated play of lab experimental games may complicate interpretation in some contexts, several factors allay this concern here. First, unlike strategic games, the "dictator game" has no second "player" about which to learn. Second, the two rounds of play were fully two years apart.

[40] We follow Bloom et al. (2015) in using the Maslach Burnout Inventory to capture job satisfaction and the Clark-Tellgen index of positive and negative affect to capture the overall attitude of teachers.

[41] We follow the phrasing used in the surveys run by Muralidharan and Sundararaman (2011a).

[42] Consistent with our failure to find "surprise effects" in student learning, there is no evidence that the re-randomization resulted in hostility toward P4P; if anything the reverse.

Table 6—Characteristics of Retained Recruits at Endline

	Grading task	Dictator game
Experienced P4P	0.68	−0.04
	(0.57)	(0.06)
Observations	170	169

Notes: For each estimated parameter, the table reports the point estimate and *p*-value in parentheses. 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 on a (raw) scale from 0 to 30; in the second column, it 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.

IV. Discussion

Compositional Margin.—To recap from Section IIIA, we find no evidence of an advertised treatment impact on the measured quality of applicants for upper-primary teaching posts in study districts, but we do find evidence of an advertised treatment impact on the measured intrinsic motivation of individuals who are placed into study schools. We draw three conclusions from these results.

First, potential applicants were aware of, and responded to, the labor market intervention. The differences in distributions across advertised treatment arms in Figure 3, panel B (dictator game share sent) and Figure 4, panel B (TVA in year 2) show that the intervention changed behavior. Since these differences are for placed recruits and not applicants, it could be that this behavior change was on the labor demand rather than supply side. In online Appendix A Figure A.2, 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. It is clear from the figure that the predicted probabilities are similar across P4P and FW labor markets. We also test formally whether the probability of hiring, as a function of CV characteristics (TTC score, age, and gender), is the same under both P4P and FW advertisements.[43] We find no statistically significant differences across advertised treatment arms.

Second, the 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, online Appendix D Table D.1 shows that the rank correlation between the baseline dictator game share sent by recruits and their TVA 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 online Appendix A Table A.4 and the distributions of TVA 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.

[43] Note that this is a sufficient but not necessary test of the absence of a demand-side response. It is sufficient because districts do not interview applicants, so CVs give us the full set of characteristics that could determine hiring. It is not necessary, however, because we observe hires rather than offers. The probability that an offer is accepted could be affected by the advertised contract associated with that post, even if applicants apply to jobs of both types and even if District Education Offices do not take contract offer types into account when selecting the individuals to whom they would like to make offers.

Finally, 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—neither in TTC scores or in the grading task scores that districts could in principle adopt.[44] This suggests that there is not an obvious demand-side policy alternative to contractually induced supply-side selection.

Effort Margin.—To recap from Section IIIB, 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 IIIC, 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 online Appendix D, we show that the rank correlation between recruits' baseline grading task IRT score and their TVA is positive. However, in Section IIIC 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 TVA. 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.

Total Effect.—The total effect of the P4P contract combines both the advertised and experienced impacts: $\tau_A + \tau_E$. By the second year of the study, the within-year total effect of P4P is $0.04 + 0.16 = 0.20$ standard deviations of pupil learning, which is statistically significant at the 1 percent level. Roughly four-fifths of the total effect can thus be attributed to increased teacher effort, while the remainder arises from supply-side selection during recruitment. At a minimum, our results suggest that in relation to positive effort-margin effects, fears of P4P causing motivational crowd out among new public sector employees may be overstated.

Our estimates raise the question of why this effect is so much stronger in year 2 compared to year 1, particularly on the effort margin. One interpretation is that this

[44] 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.

is because it takes time for recruits to settle into 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 estimates of longer-term impacts.

V. 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 P4P. 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.

Given these encouraging results, it is natural to ask whether it would be feasible and cost effective to implement this P4P contract at scale. We worked closely with the government to design a contract that was contextually feasible and well grounded in theory. A composite P4P metric was used to avoid narrowly emphasizing any single aspect of teacher performance, and when measuring learning, we followed the pay-for-percentile approach that aims to give all teachers a fair chance regardless of the composition of the students they teach. We also took care to ensure that the P4P contract, if successful, could be built into the growth path of teacher wages. While a larger bonus might have elicited stronger impacts, the expected value of the P4P bonus was set at 3 percent of teacher salaries to be commensurate with annual teacher salary increments (and discretionary pay in other sectors under Rwanda's *imihigo* system of performance contracts for civil servants).

The fact that we compared a P4P contract with an expenditure-equivalent FW alternative that is equal in magnitude to annual teacher salary increments means that it is reasonable to think about cost effectiveness primarily in terms of measurement. For pupil learning, the minimum requirement for the P4P contract we study is a system of repeated annual assessments across grades and key subjects.[45] Measurement of the other aspects of performance—teacher presence, preparation, and pedagogy—can in principle be conducted by head teachers or district staff (who are increasingly being asked to monitor teacher performance) at modest cost.

There are nonetheless limitations of our work. Inasmuch as the impacts on either the compositional or effort margin might differ after five or ten years, there is certainly scope for further study of this topic in low- and middle-income countries. For instance, it would be interesting to explore whether long-term P4P commitments influence early-career decisions to train as a teacher; our study restricts attention to employment choices by individuals who have already received TTC degrees.

[45] Following a 2019 Cabinet Resolution, Rwanda is establishing a Comprehensive Assessment that is intended to serve this purpose.

Another set of issues relate to unintended consequences of P4P. We found that advertised P4P attracted teachers with lower intrinsic motivation, as measured by the share sent in the framed baseline dictator game. It is possible that the students taught by these more self-regarding teachers became more self-regarding themselves or otherwise developed different soft skills. We also found that experienced P4P improved performance on three of the four incentivized dimensions of the composite metric: teacher presence and pedagogy and pupil learning. It is conceivable that the students in P4P schools may have been impacted by "multitasking" as teachers focused on these dimensions to the detriment of others. Since we did not measure aspects of student development beyond test score gains, it would be interesting to explore these issues in future work.

Rwanda's labor market has a characteristic that is 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 positive effects we find in Rwanda of a multidimensional, pay-for-percentile contract—improving performance without dampening employee satisfaction—will generalize to settings where public sector wage premiums differ remains an open question, for the education sector and beyond.

REFERENCES

Adnot, Melinda, Thomas Dee, Veronica Katz, and James Wyckoff. 2017. "Teacher Turnover, Teacher Quality, and Student Achievement in DCPS." *Educational Evaluation and Policy Analysis* 39 (1): 54–76.

Anderson, Michael L., and Jeremy Magruder. 2017. "Split-Sample Strategies for Avoiding False Discoveries." NBER Working Paper 23544.

Ashraf, Nava, Oriana Bandiera, Edward Davenport, and Scott S. Lee. 2020. "Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the Delivery of Public Services." *American Economic Review* 110 (5): 1355–94.

Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack. 2014. "No Margin, No Mission? A Field Experiment on Incentives for Public Service Delivery." *Journal of Public Economics* 120: 1–17.

Ashraf, Nava, James Berry, and Jesse M. Shapiro. 2010. "Can Higher Prices Stimulate Product Use? Evidence from a Field Experiment in Zambia." *American Economic Review* 100 (5): 2383–413.

Banuri, Sheheryar, and Philip Keefer. 2016. "Pro-Social Motivation, Effort and the Call to Public Service." *European Economic Review* 83: 139–64.

Barlevy, Gadi, and Derek Neal. 2012. "Pay for Percentile." *American Economic Review* 102 (5): 1805–31.

Bénabou, Roland, and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies* 70 (3): 489–520.

Biasi, Barbara. 2019. "The Labor Market for Teachers Under Different Pay Schemes." NBER Working Paper 24813.

Bloom, Nicholas, James Liang, John Roberts, and Zhichun Jenny Ying. 2015. "Does Working from Home Work? Evidence from a Chinese Experiment." *Quarterly Journal of Economics* 130 (1): 165–218.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, and Christophe Rockmore. 2017. "Enrollment without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa." *Journal of Economic Perspectives* 31 (4): 185–204.

Brock, J. Michelle, Andreas Lange, and Kenneth L. Leonard. 2016. "Generosity and Prosocial Behavior in Healthcare Provision: Evidence from the Laboratory and Field." *Journal of Human Resources* 51 (1): 133–62.

Callen, Michael, Saad Gulzar, Ali Hasanain, Muhammad Khan, and Arman Rezaee. 2020. "Data and Policy Decisions: Experimental Evidence from Pakistan." *Journal of Development Economics* 146: 102523.

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

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104 (9): 2593–632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104 (9): 2633–79.

Chingos, Matthew M., and Martin R. West. 2012. "Do More Effective Teachers Earn More Outside the Classroom?" *Education Finance and Policy* 7 (1): 8–43.

Cohen, Jessica, and Pascaline Dupas. 2010. "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment." *Quarterly Journal of Economics* 125 (1): 1–45.

Dal Bó, Ernesto, and Frederico Finan. 2016. "At the Intersection: A Review of Institutions in Economic Development." EDI Working Paper 16/11.01.

Dal Bó, Ernesto, Frederico Finan, and Martin A. Rossi. 2013. "Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service." *Quarterly Journal of Economics* 128 (3): 1169–218.

Danielson, Charlotte. 2007. *Enhancing Professional Practice: A Framework for Teaching.* 2nd ed. Alexandria, VA: Association for Supervision and Curriculum Development.

Deci, Edward L., and Richard M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior*. New York: Plenum.

Dee, Thomas S., and James Wyckoff. 2015. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT." *Journal of Policy Analysis and Management* 34 (2): 267–97.

Delfgaauw, Josse, and Robert Dur. 2008. "Incentives and Workers' Motivation in the Public Sector." *Economic Journal* 118 (525): 171–91.

Deserranno, Erika. 2019. "Financial Incentives as Signals: Experimental Evidence from the Recruitment of Village Promoters in Uganda." *American Economic Journal: Applied Economics* 11 (1): 277–317.

DiCiccio, Cyrus J., and Joseph P. Romano. 2017. "Robust Permutation Tests for Correlation and Regression Coefficients." *Journal of the American Statistical Association* 112 (519): 1211–20.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4): 1241–78.

Eckel, Catherine C., and Philip J. Grossman. 1996. "Altruism in Anonymous Dictator Games." *Games and Economic Behavior* 16 (2): 181–91.

Einav, Liran, Amy Finkelstein, Stephen P. Ryan, Paul Schrimpf, and Mark R. Cullen. 2013. "Selection on Moral Hazard in Health Insurance." *American Economic Review* 103 (1): 178–219.

Fafchamps, Marcel, and Julien Labonne. 2017. "Using Split Samples to Improve Inference on Causal Effects." *Political Analysis* 25: 465–82.

Finan, Frederico, Benjamin A. Olken, and Rohini Pande. 2017. "The Personnel Economics of the State." In *Handbook of Field Experiments*. Vol. 2, edited by Abhijit Banerjee and Esther Duflo, 467–514. Amsterdam: Elsevier.

Gilligan, Daniel O., Naureen Karachiwalla, Ibrahim Kasirye, Adrienne Lucas, and Derek A. Neal. Forthcoming. "Educator Incentives and Educational Triage in Rural Primary Schools." *Journal of Human Resources*.

Glewwe, Paul, and Karthik Muralidharan. 2016. "Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." In *Handbook of the Economics of Education*. Vol. 5, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 653–743. Amsterdam: Elsevier.

Hanushek, Eric A., and Ludger Woessmann. 2012. "Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation." *Journal of Economic Growth* 17 (4): 267–321.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21 (1): 1–20.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge, UK: Cambridge University Press.

Imberman, Scott A. 2015. "How Effective Are Financial Incentives for Teachers?" *IZA World of Labor* 158.

Jackson, C. Kirabo, Jonah E. Rockoff, and Douglas O. Staiger. 2014. "Teacher Effects and Teacher-Related Policies." In *Annual Review of Economics*. Vol. 6, edited by Kenneth J. Arrow and Timothy F. Bresnahan, 801–25. Palo Alto, CA: Annual Reviews.

Karlan, Dean, and Jonathan Zinman. 2009. "Observing Unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment." *Econometrica* 77 (6): 1993–2008.

Krepps, David. 1997. "Intrinsic Motivation and Extrinsic Incentives." *American Economic Review* 87 (2): 359–64.

Lazear, Edward P. 2000. "Performance Pay and Productivity." *American Economic Review* 90 (5): 1346–61.

Lazear, Edward P. 2003. "Teacher Incentives." *Swedish Economic Policy Review* 10 (3): 179–214.

Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin. 2018. "Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools." AEA RCT Registry. https://doi.org/10.1257/rct.2565-5.0.

Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin. 2021. "Replication data for: Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools." American Economic Association [publisher]. Inter-university Consortium for Political and Social Research [distributor]. doi: 10.3886/E121941V1.

Loyalka, Prashant, Sean Sylvia, Chengfang Liu, James Chu, and Yaojiang Shi. 2019. "Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement." *Journal of Labor Economics* 37 (3): 621–62.

Mbiti, Isaac, Mauricio Romero, and Youdi Schipper. 2019. "Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania." NBER Working Paper 25903.

Muralidharan, Karthik, Mauricio Romero, and Kaspar Wüthrich. 2019. "Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments." NBER Working Paper 26562.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011a. "Teacher Opinions on Performance Pay: Evidence from India." *Economics of Education Review* 30 (3): 394–403.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011b. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119 (1): 39–77.

National Institute of Statistics of Rwanda. 2017. "Labour Force Survey." Republic of Rwanda. https://microdata.statistics.gov.rw/index.php/catalog/81 (accessed December 1, 2019).

Neal, Derek A. 2011. "The Design of Performance Pay in Education." In *Handbook of the Economics of Education*. Vol. 4, edited by Eric A. Hanushek, Stephen J. Machin, and Ludger Woessmann, 495–550. Amsterdam: North Holland.

Olken, Benjamin A. 2015. "Promises and Perils of Pre-Analysis Plans." *Journal of Economic Perspectives* 29 (3): 61–80.

Rothstein, Jesse. 2015. "Teacher Quality Policy When Supply Matters." *American Economic Review* 105 (1): 100–130.

Sojourner, Aaron J., Elton Mykerezi, and Kristine L. West. 2014. "Teacher Pay Reform and Productivity: Panel Data Evidence from Adoptions of Q-Comp in Minnesota." *Journal of Human Resources* 49 (4): 945–81.

Staiger, Douglas O., and Jonah E. Rockoff. 2010. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24 (3): 97–118.

Stecher, Brian M., Deborah J. Holtzman, Michael S. Garet, Laura S. Hamilton, John Engberg, Elizabeth D. Steiner, Abby Robyn, et al. 2018. *Improving Teacher Effectiveness: Final Report: The Intensive Partnerships for Effective Teaching Through 2015–2016*. Santa Monica, CA: RAND Corporation.

Zeitlin, Andrew. 2021. "Teacher Turnover in Rwanda." *Journal of African Economies* 30 (1): 81–102.